Introduction to Data Management PROJECT REPORT

(Project Semester August-December 2021)

PROJECT REPORT

ON

Goodreadsbooks data

Submitted by

BACHU VARUN

11912352

Program: Bachelor of Technology

Section: KM006

Course Code: INT217

Under the Guidance of

Komal Arora: 17783

Assistant Professor

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

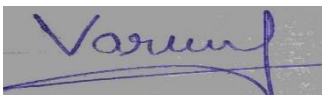
DECLARATION

I, Bachu Varun, student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 18/12/2021

Bachu Varun

Registration No: 11912352

Signature: 

ACKNOWLEDGEMENT

Primarily I'd thank God for being able to complete my project with success. Then I'd like to thank my mentor ***Ms. Komal Arora****, whose valuable guidance has been the ones that helped me patch this project and make it full proof success in contribution towards the completion of this project.*

Finally, I'd rather thanks to ***Lovely Professional University,*** *and my parent's inspiration, who gave me this golden opportunity to learn many new things, to learn another aspect of life.*

-Bachu Varun

CONTENTS:

Sr No.	***Title***	***Page No.***
1	*Introduction*	*5*
2	*Objectives/Scope of the Analysis*	*6*
3	*Source of dataset*	*7-8*
4	*ETL Process*	*9-21*
5	*Analysis of dataset*	*22-31*
6	*List of Analysis with results*	*32-36*
7	*Final Dashboard*	*37*
8	*Bibliography*	*38*

INTRODUCTION

Goodreadsbooks is a data of some good books. It is very useful to the book-lovers out there, irrespective of language and publication and all of that. We can use it to our liking, find what book we should be reading next. This data consists of bookID, Title, Authors, Average_rating, isbn13(International Standard Book Number), language_code, number of pages, ratings_count, text_reviews_count, publication_date, publisher.

Tableau is a data analysis and visualization tool which is commonly used in today's industry. Many organizations still find it important for the research relevant to data science. The ease ofuse of Tableau is due to it providing a drag and drop interface. This feature helps to perform tasks like sorting, comparing, and analyzing, very easily and fast. Tableau is also compatible with multiple sources, including Excel, SQL Server, and cloud-based data repositories which makes it an excellent choice for Data Scientists.

OBJECTIVES/SCOPE OF ANALYSIS

After analysis of the dataset, the aim of this project is to give answer of given objectives in easy way:

- *Total books published in each year*
- *Top 50 average rated books*
- *Most rating counts containing books in each year*
- *No. of books written in each language*
- *Top 10 text reviewed books from whole dataset*

SOURCE OF DATASET:

Source of dataset: https://www.kaggle.com/jealousleopard/goodreadsbooks

Kaggle is an online community for data scientists and machine learners, developed by Google. Kaggle allows users to find and publish data sets, explore, and build models in a web- based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud- based workbench for data science, and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

This data science project analyzes the Goodreadsbooks dataset. It was created for the (B. Tech CSE fifth semester Introduction to Data Science course) project.

Every part of the dataset consists of multiple of punctuation errors which is cleaned in the ETL process.

Sample of dataset with data fields is given below:

Here we can see multiple columns with names such as BookID, Title, Authors, Average rating, isbn13, language code, number of pages, text reviews count, publication date, publisher…so on

	A	B	C	D	E	F	G	H	I	J	K
1	bookID	title	authors	average_rating	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
2	1	Harry Potter and the Half-Blo	J.K. Rowling/Mary GrandPré	4.57	9780439785969	eng	652	2095690	27591	16-09-06	Scholastic Inc.
3	2	Harry Potter and the Order of	J.K. Rowling/Mary GrandPré	4.49	9780439358071	eng	870	2153167	29221	01-09-04	Scholastic Inc.
4	3	Harry Potter and the Chambe	J.K. Rowling	4.42	9780439554893	eng	352	6333	244	01-11-03	Scholastic
5	4	Harry Potter and the Prisoner	J.K. Rowling/Mary GrandPré	4.56	9780439655484	eng	435	2339585	36325	01-05-04	Scholastic Inc.
6	5	Harry Potter Boxed Set Books	J.K. Rowling/Mary GrandPré	4.78	9780439682589	eng	2690	41428	164	13-09-04	Scholastic
7	6	Unauthorized Harry Potter Bo	W. Frederick Zimmerman	3.74	9780976540601	en-US	152	19	1	26-04-05	Nimble Books
8	7	Harry Potter Collection (Harr	J.K. Rowling	4.73	9780439827607	eng	3342	28242	808	12-09-05	Scholastic
9	8	The Ultimate Hitchhiker's Gui	Douglas Adams	4.38	9780517226957	eng	815	3628	254	01-11-05	Gramercy Books
10	9	The Ultimate Hitchhiker's Gui	Douglas Adams	4.38	9780345453747	eng	815	249558	4080	30-04-02	Del Rey Books
11	10	The Hitchhiker's Guide to the	Douglas Adams	4.22	9781400052929	eng	215	4930	460	03-08-04	Crown
12	11	The Hitchhiker's Guide to the	Douglas Adams/Stephen Fry	4.22	9780739322208	eng	6	1266	253	23-03-05	Random House Audio
13	12	The Ultimate Hitchhiker's Gui	Douglas Adams	4.38	9780517149256	eng	815	2877	195	17-01-96	Wings Books
14	13	A Short History of Nearly Ever	Bill Bryson	4.21	9780767908184	eng	544	248558	9396	14-09-04	Broadway Books
15	14	Bill Bryson's African Diary	Bill Bryson	3.44	9780767915069	eng	55	7270	499	03-12-02	Broadway Books
16	15	Bryson's Dictionary of Troubl	Bill Bryson	3.87	9780767910439	eng	256	2088	131	14-09-04	Broadway Books
17	16	In a Sunburned Country	Bill Bryson	4.07	9780767903868	eng	335	72451	4245	15-05-01	Broadway Books
18	17	I'm a Stranger Here Myself: N	Bill Bryson	3.9	9780767903820	eng	304	49240	2211	28-06-00	Broadway Books
19	18	The Lost Continent: Travels in	Bill Bryson	3.83	9780060920081	eng	299	45712	2257	28-08-90	William Morrow Paperbacks
20	19	Neither Here nor There: Trave	Bill Bryson	3.86	9780380713806	eng	254	48701	2238	28-03-93	William Morrow Paperbacks
21	20	Notes from a Small Island	Bill Bryson	3.91	9780380727506	eng	324	80609	3301	28-05-97	William Morrow Paperbacks
22	21	The Mother Tongue: English a	Bill Bryson	3.93	9780380715435	eng	270	28489	2085	28-09-91	William Morrow Paperbacks
23	22	J.R.R. Tolkien 4-Book Boxed Se	J.R.R. Tolkien	4.59	9780345538376	eng	1728	101233	1550	25-09-12	Ballantine Books
24	23	The Lord of the Rings (The Lor	J.R.R. Tolkien	4.5	9780618517657	eng	1184	1710	91	21-10-04	Houghton Mifflin Harcourt
25	24	The Fellowship of the Ring (Tl	J.R.R. Tolkien	4.36	9780618346257	eng	398	2128944	13670	05-09-03	Houghton Mifflin Harcourt
26	25	The Lord of the Rings (The Lor	J.R.R. Tolkien/Alan Lee	4.5	9780618260584	en-US	1216	1618	140	01-10-02	Houghton Mifflin Harcourt
27	26	The Lord of the Rings: Weapo	Chris Smith/Christopher Lee/Ric	4.53	9780618391004	eng	218	19822	46	05-11-03	Houghton Mifflin Harcourt
28	27	The Lord of the Rings: Comple	Jude Fisher	4.5	9780618510825	eng	224	359	6	15-11-04	Houghton Mifflin Harcourt
29	28	Agile Web Development with	Dave Thomas/David Heinemeier H	3.84	9780976694007	eng	558	1430	59	28-07-05	Pragmatic Bookshelf
30	29	Hatchet (Brian's Saga #1)	Gary Paulsen	3.72	9780689840920	eng	208	270244	12017	01-04-00	Atheneum Books for Young Readers: R

ETL PROCESS:

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the datainto the Data Warehouse system. ETL full form is Extract, Transform and Load.

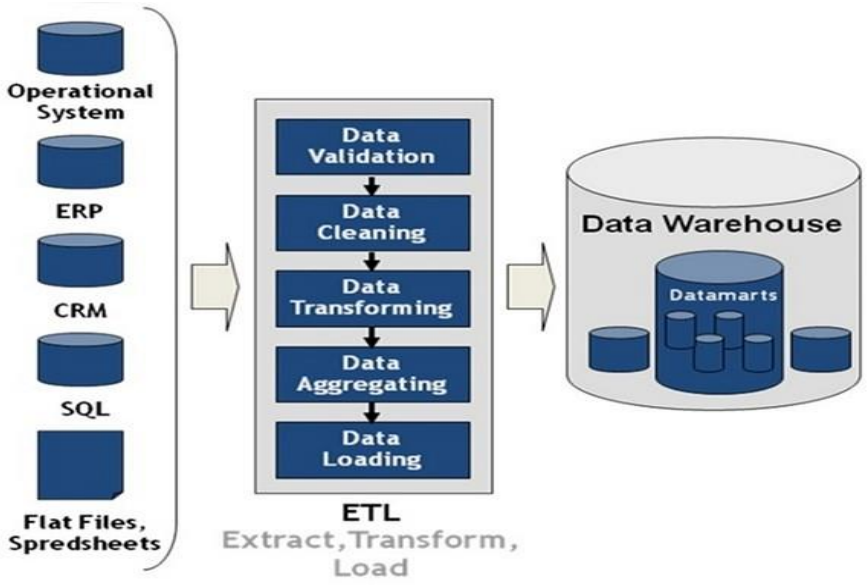
It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

Need of ETL Process

- *ETLprocess allows sample data comparison between the source and the target system.*
- *ETL is a predefined process for accessing andmanipulating source data into the targetdatabase.*
- *Allow verification of data transformation, aggregation, andcalculations rules.*

When it comes to the implementation of the ETL process, the itinerary of tasks can be divvied upinto the full form of its acronym.

1. **E – Extraction**
2. **T – Transformation**
3. **L – Loading**

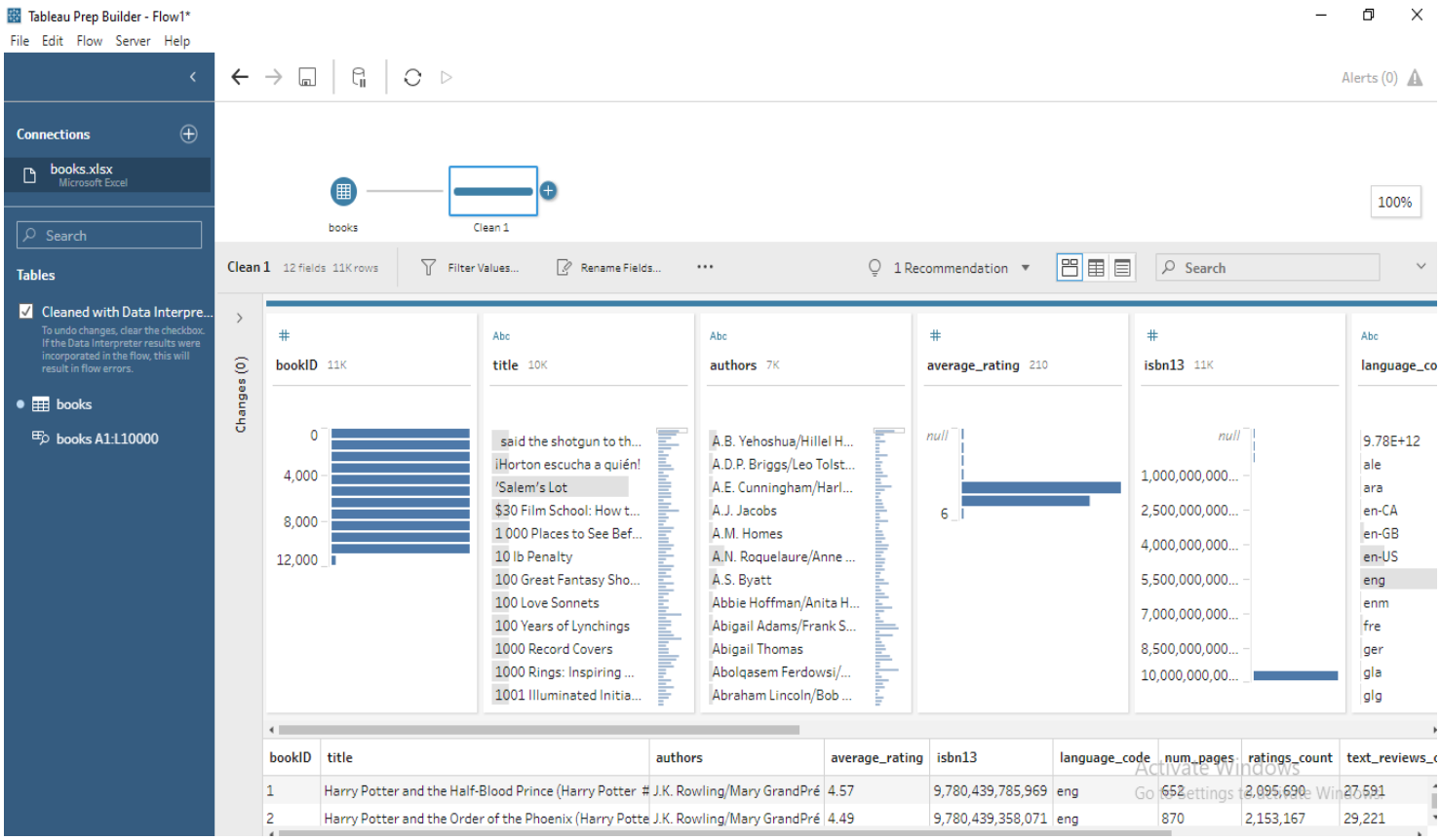


ETL Process used in Project

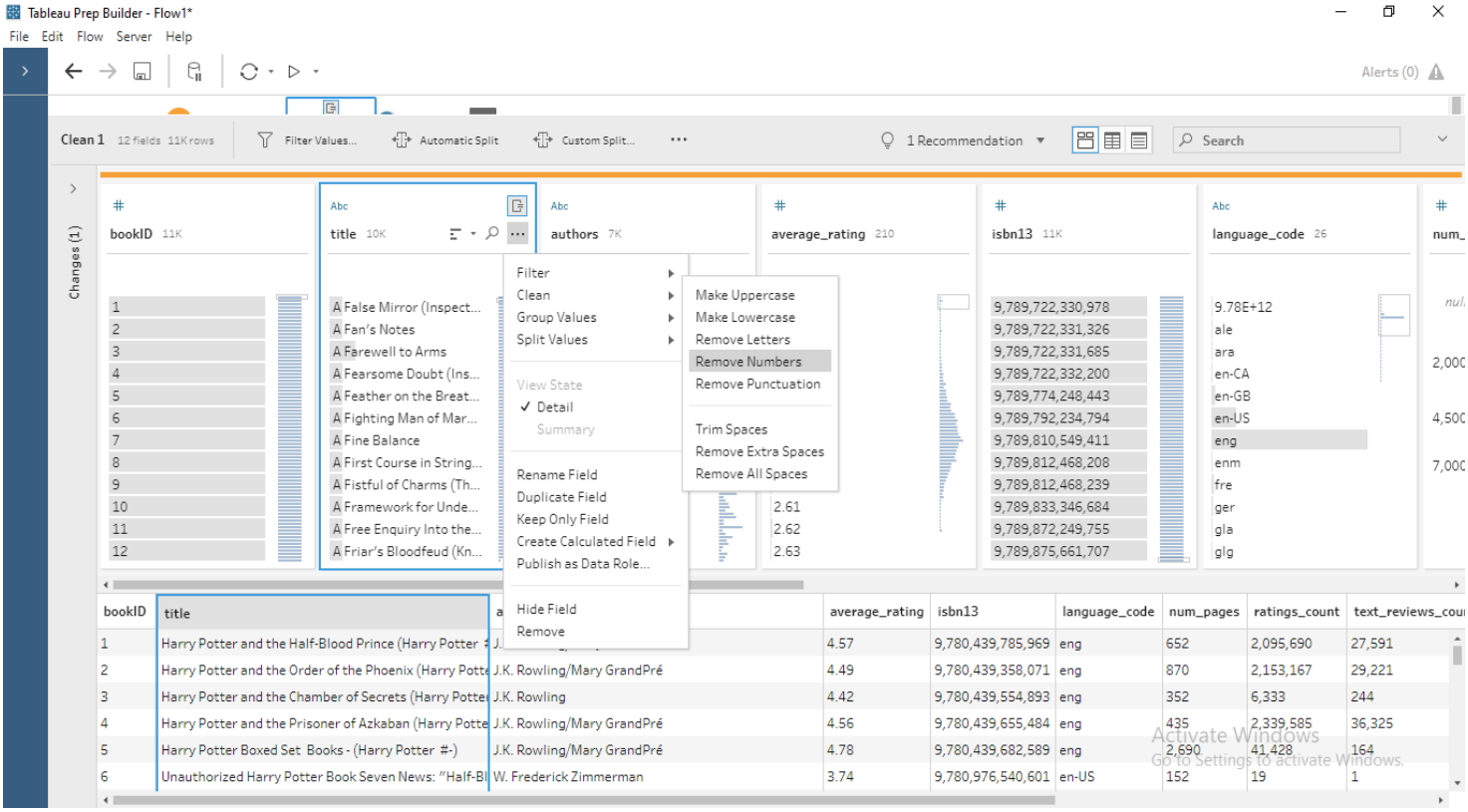
Extraction

Extracting the dataset from PC to Tableau for removing the unwanted characters, fields,spelling errors etc.

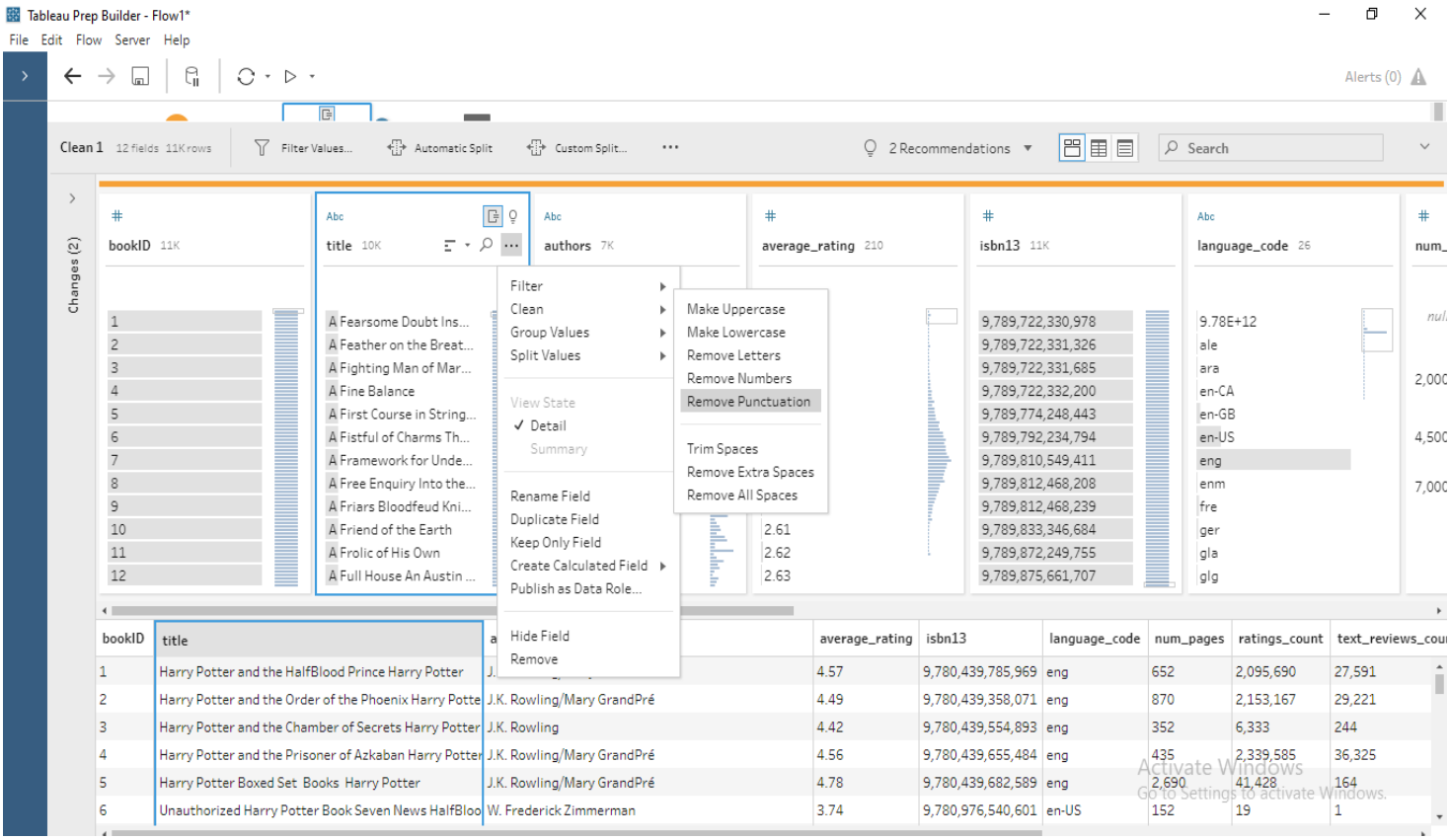
Step 1: *Opening the dataset in single table.*



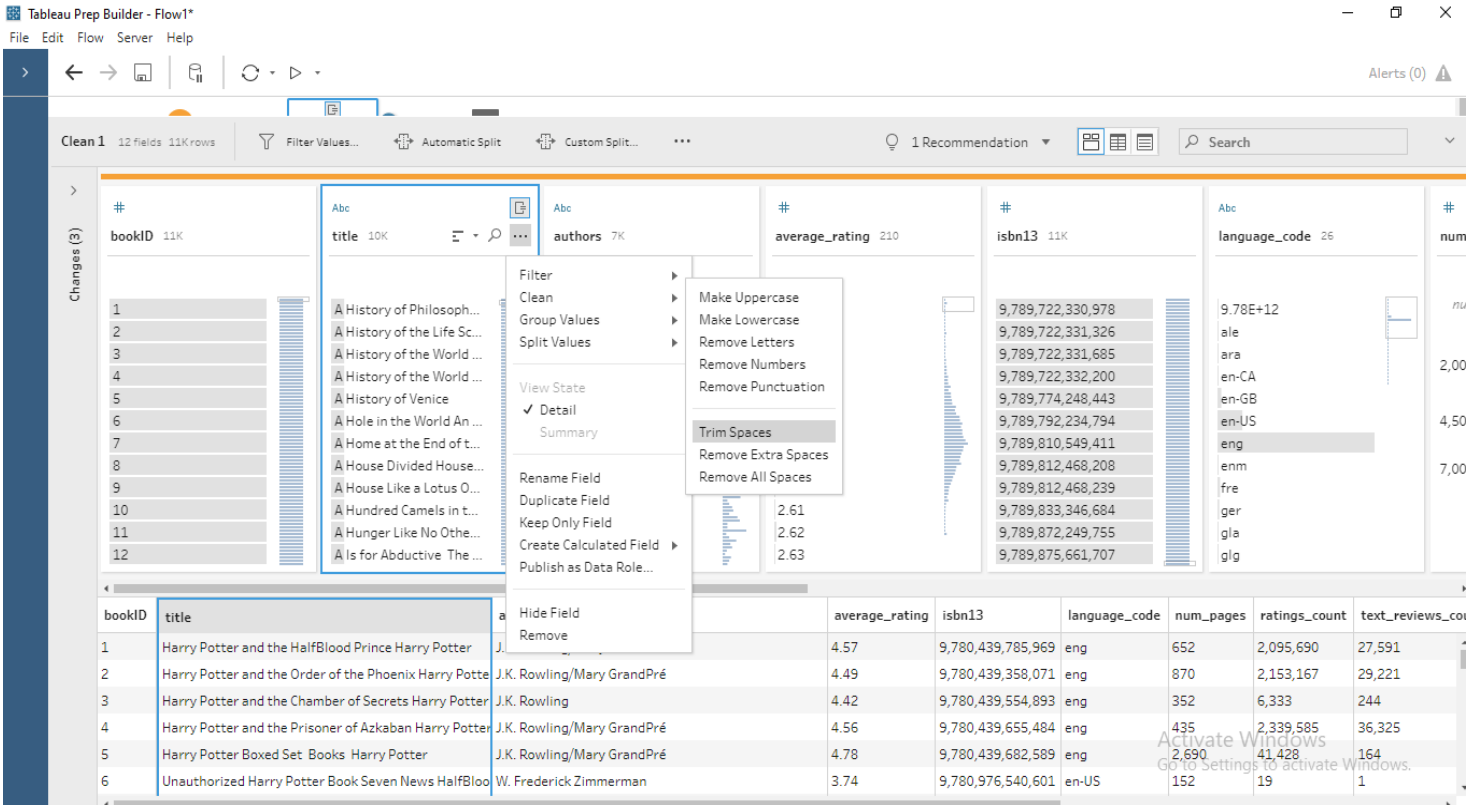
Step 2: *Removal of numbers in Title column*



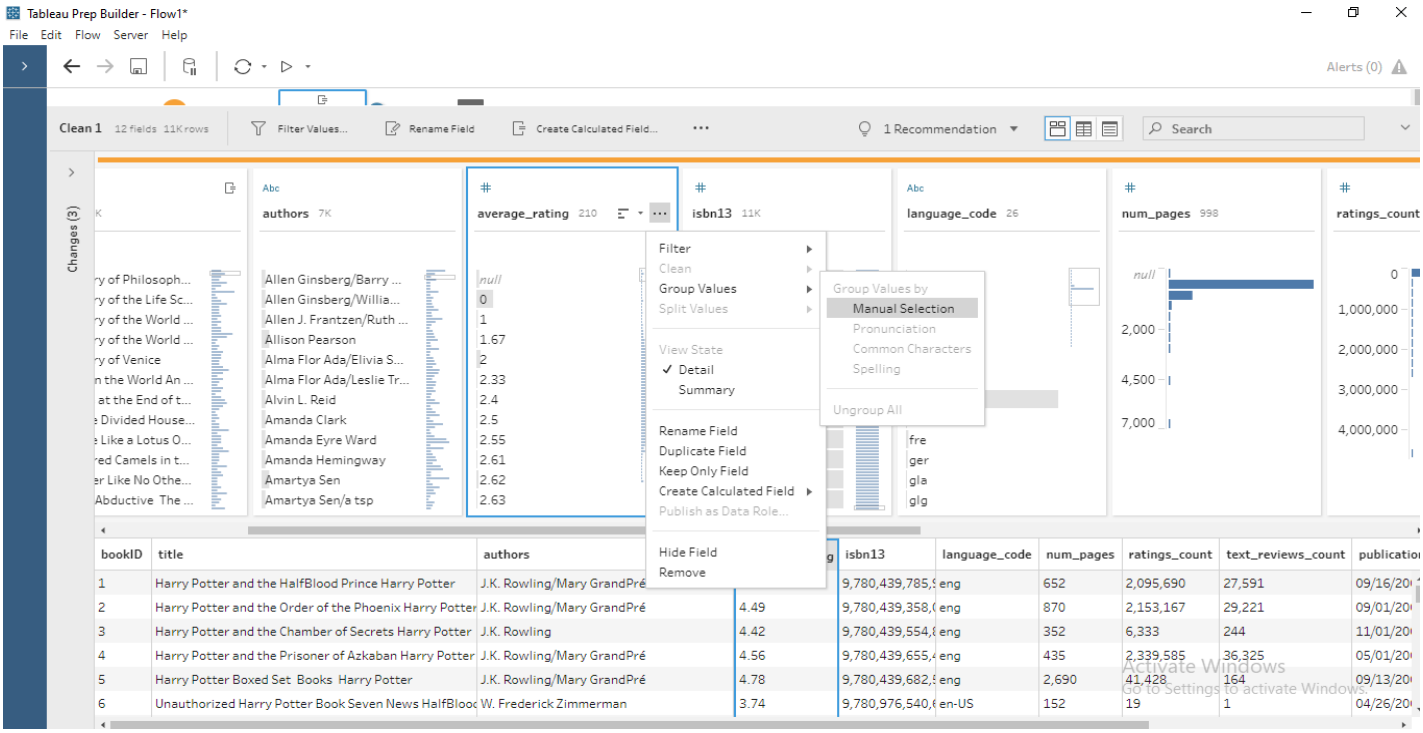
Step 3: *Removal of punctuations in title column*



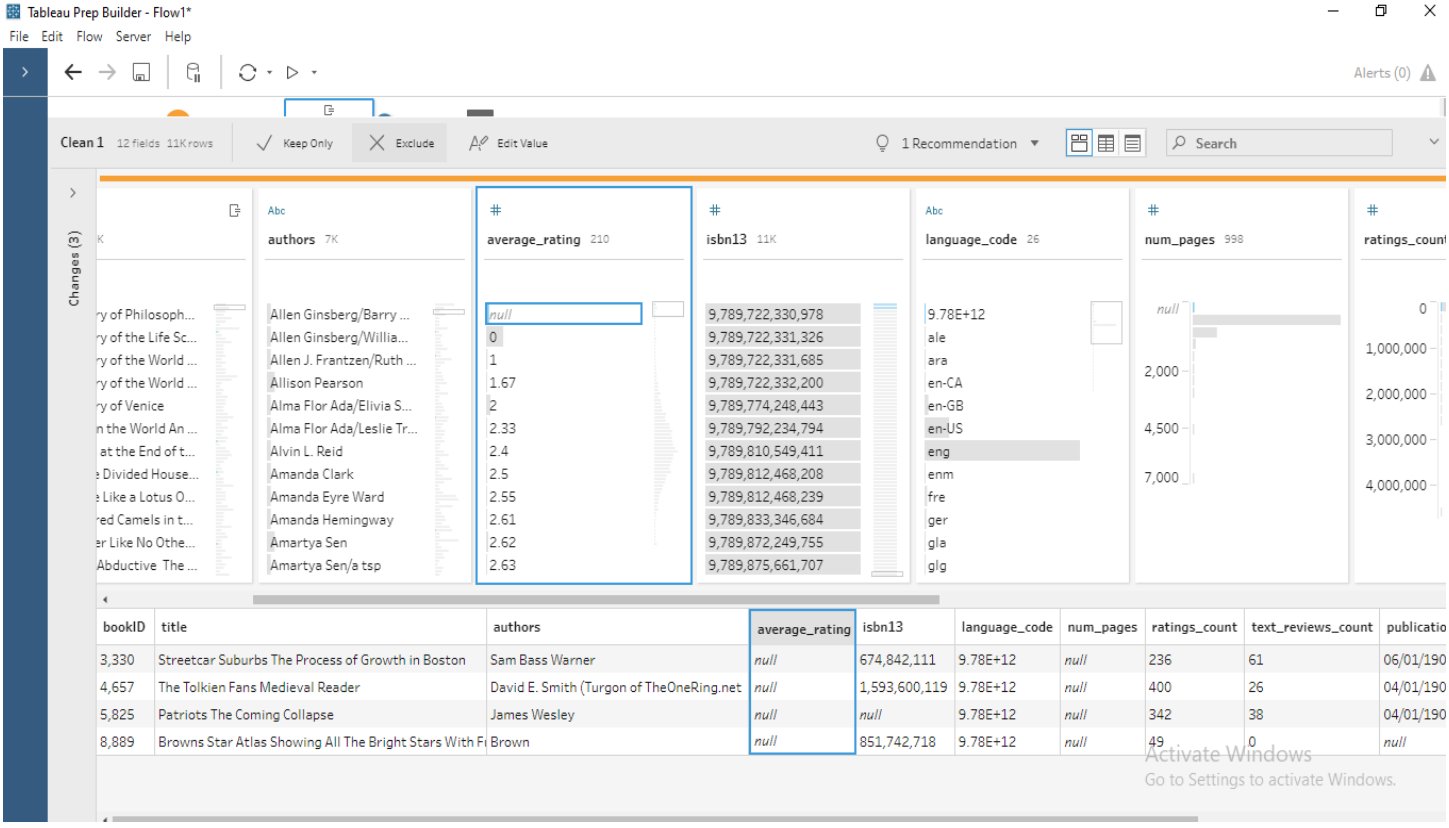
Step 4: *Trimming spaces in title column*



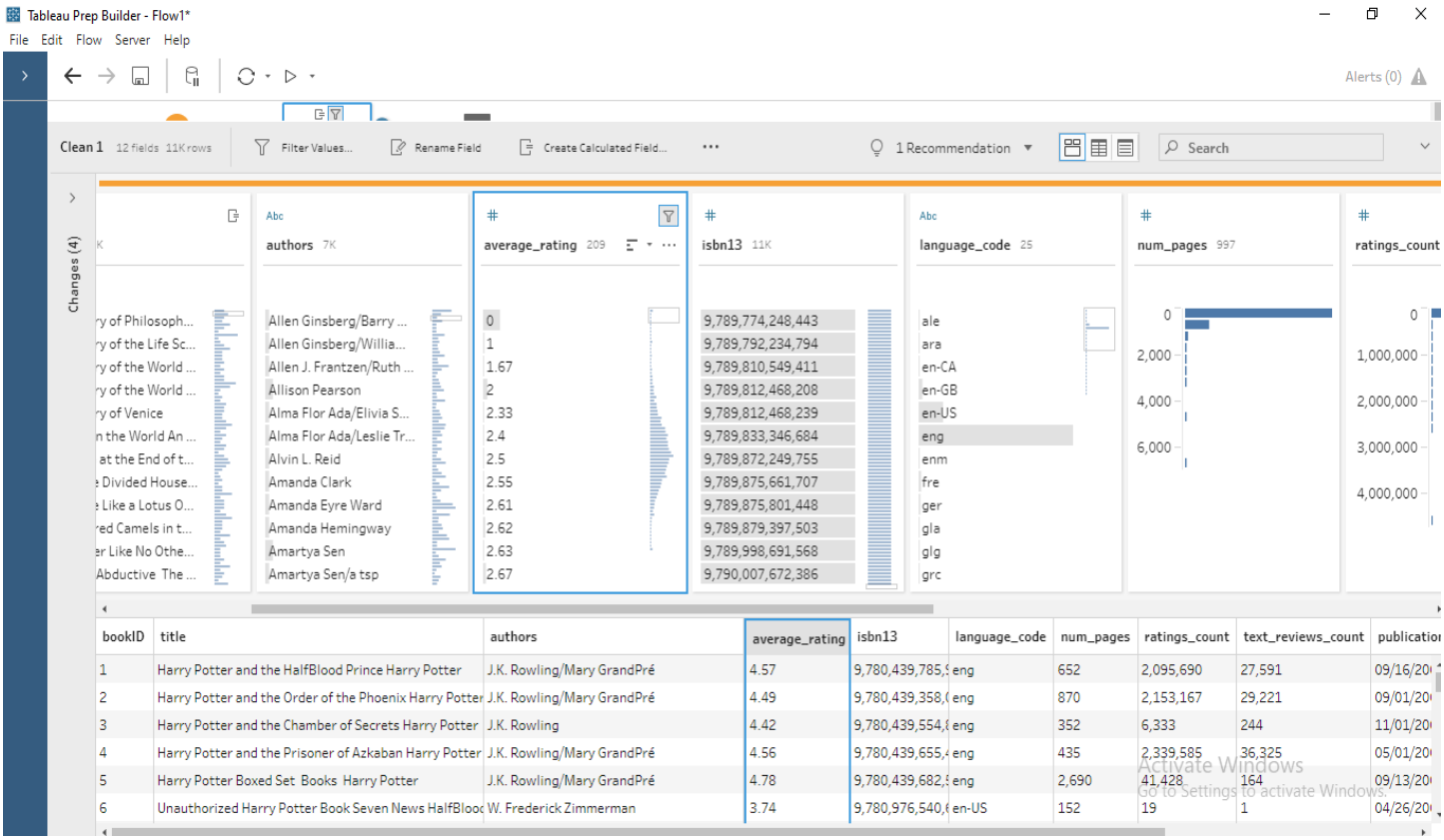
Step 5: *Removal of null values in average_rating column*



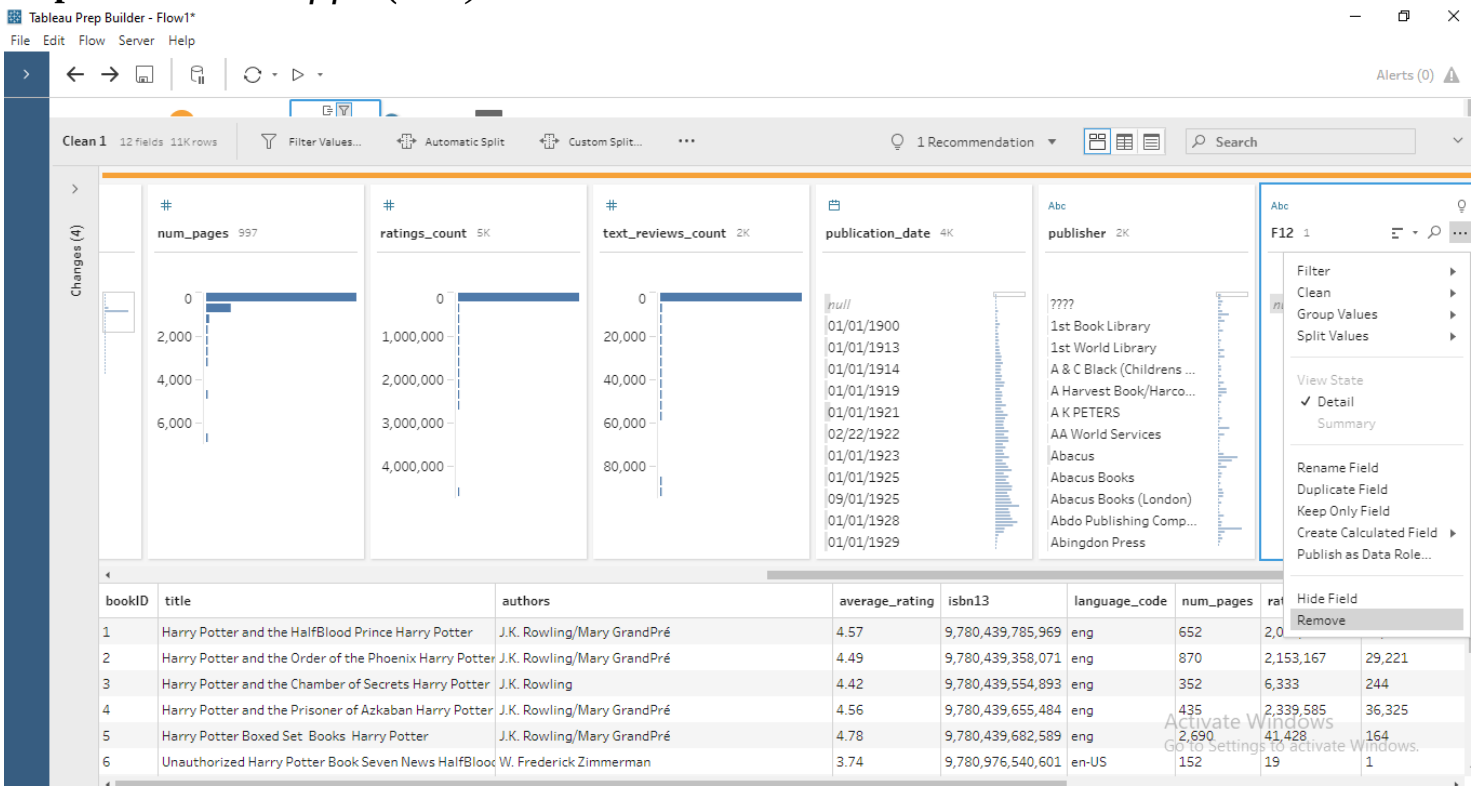
Selecting null values by manual selection and excluding it



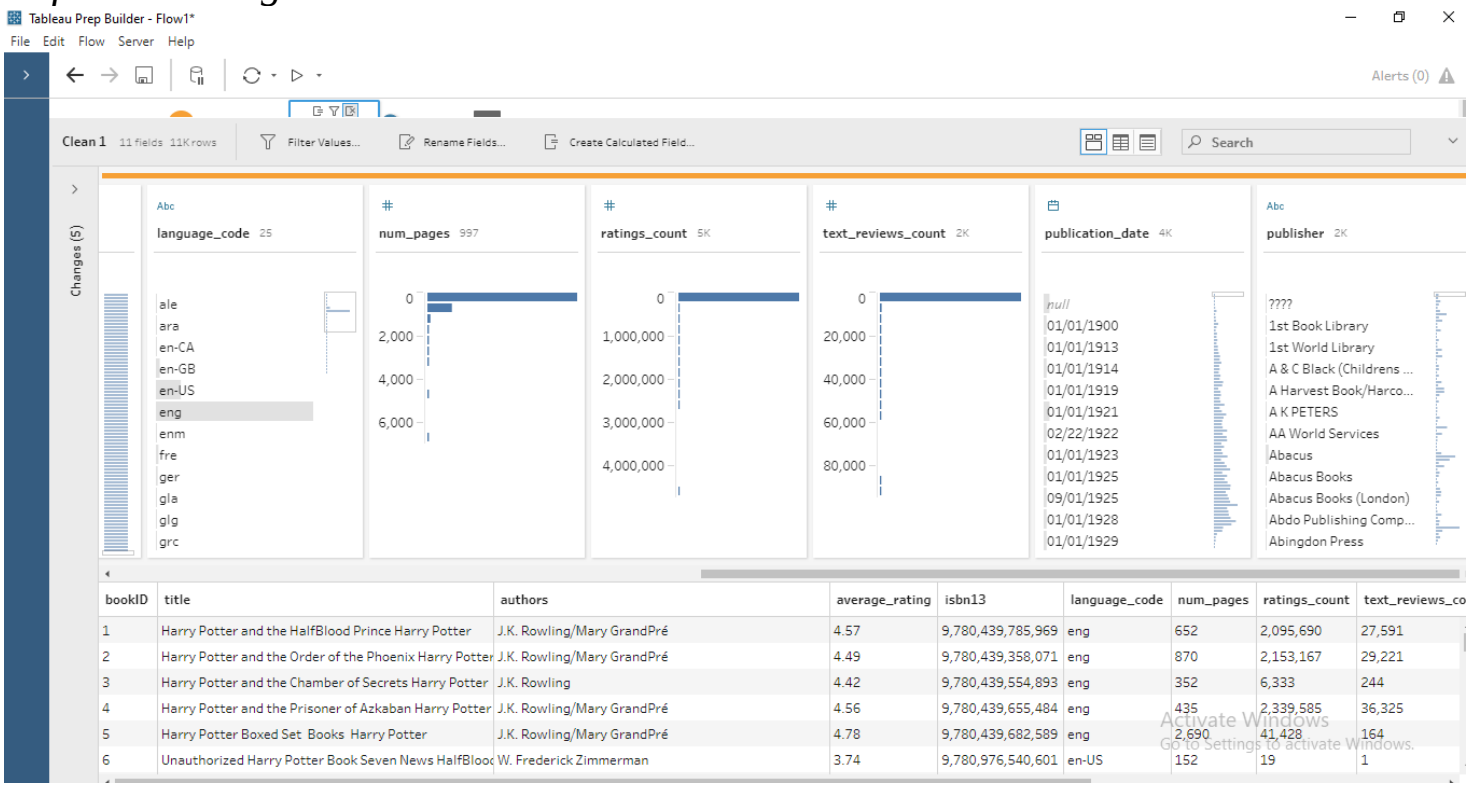
After removing null values



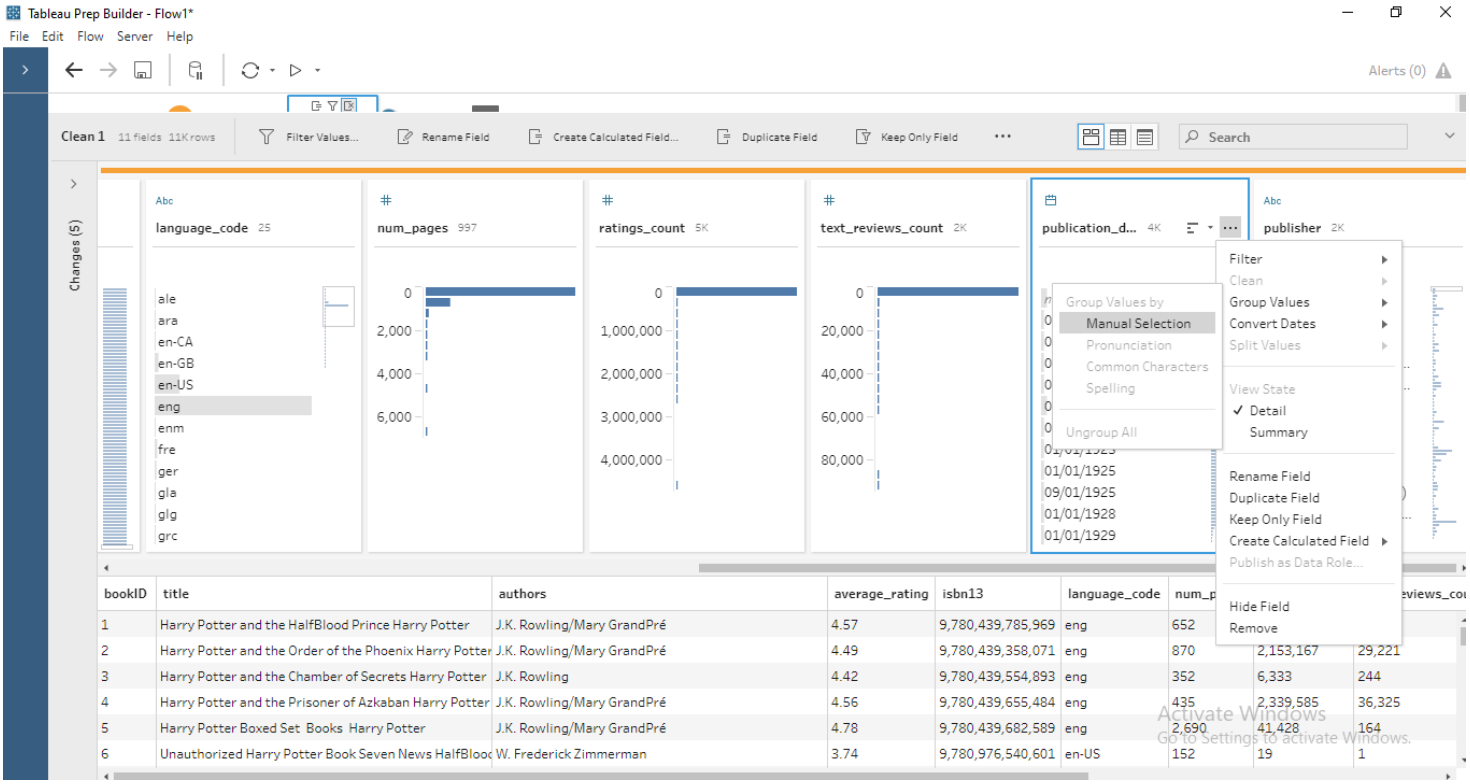
Step 6: *Removal of f12(null) column*



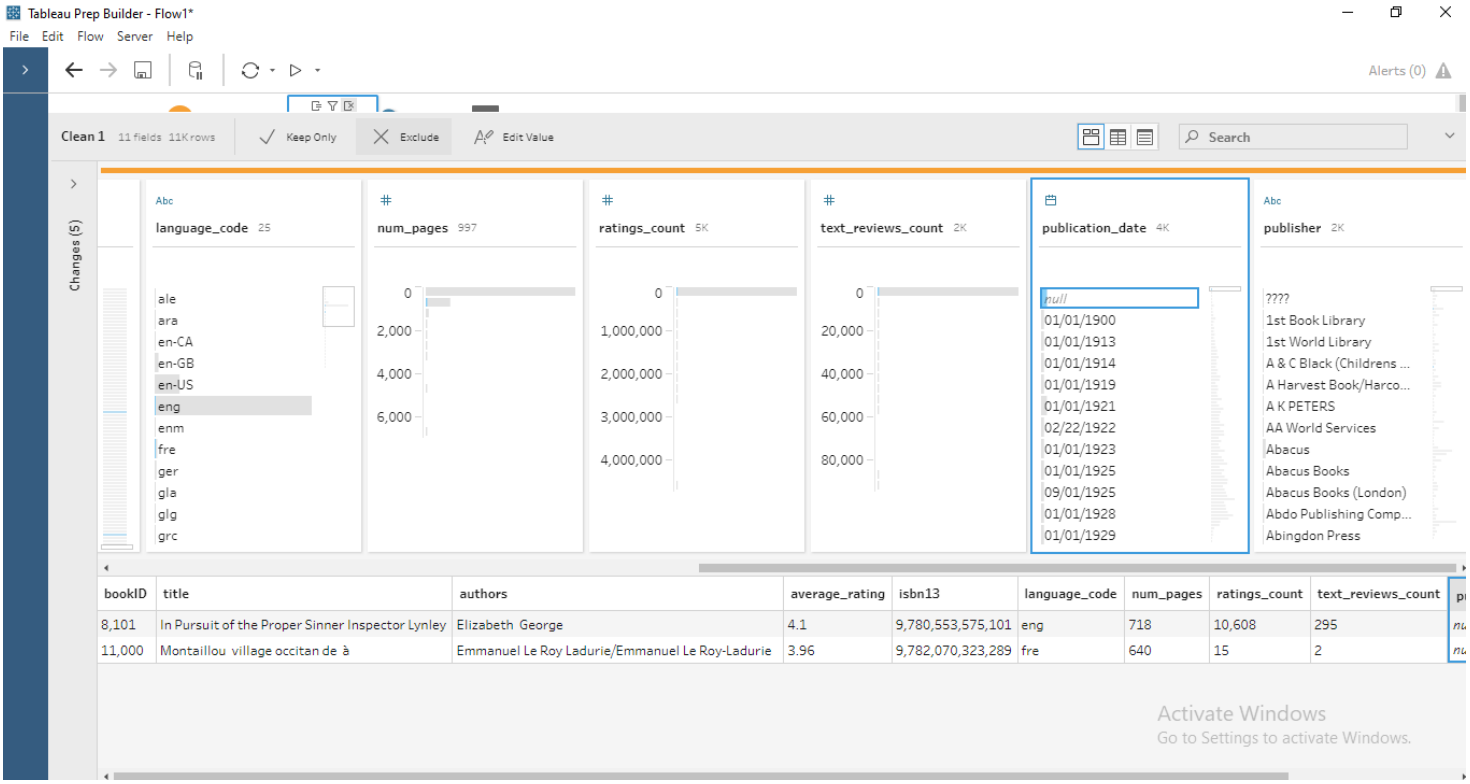
After removing column:



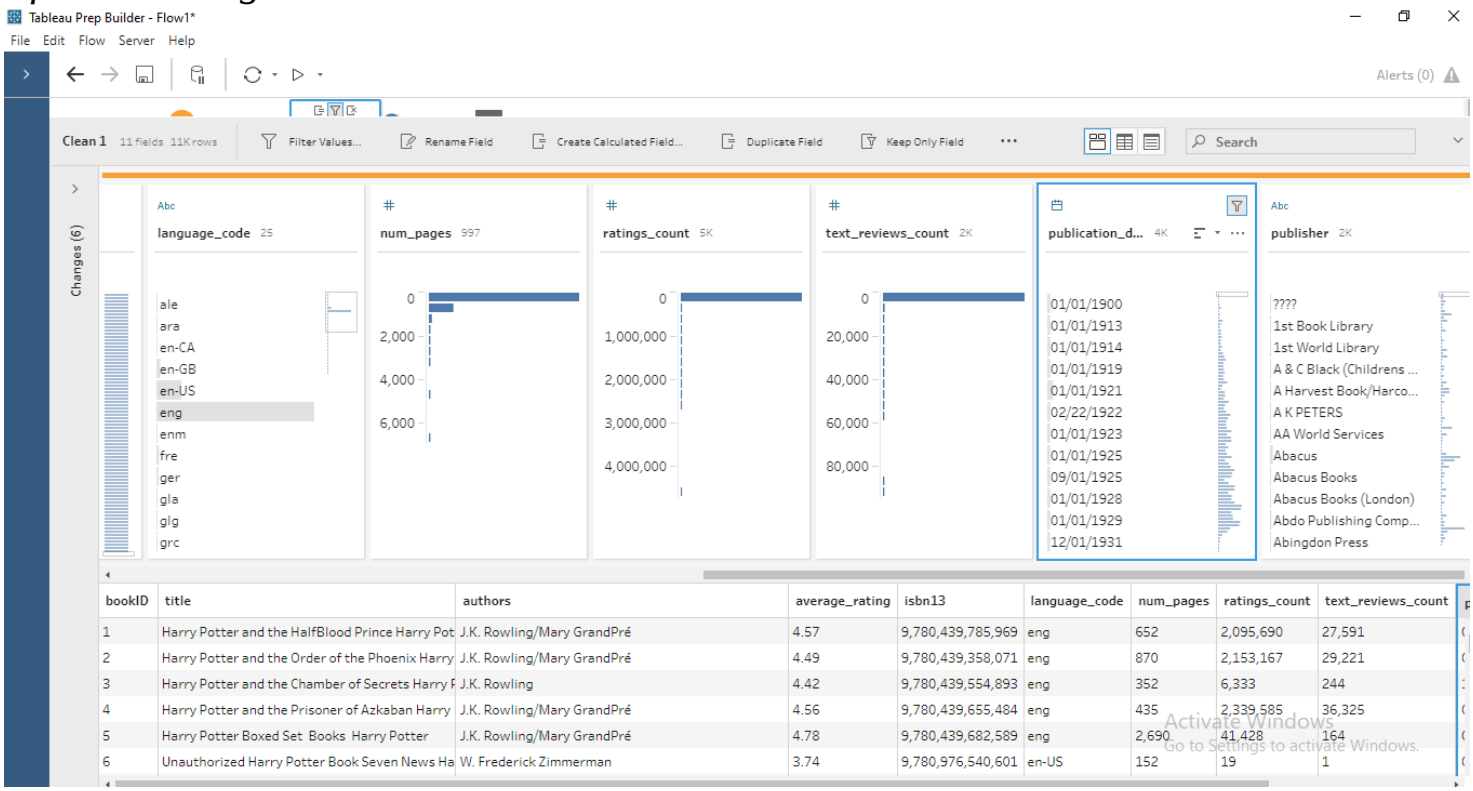
Step-7: *Removal of null values from publication_date*



Selecting null values by manual selection and excluding it



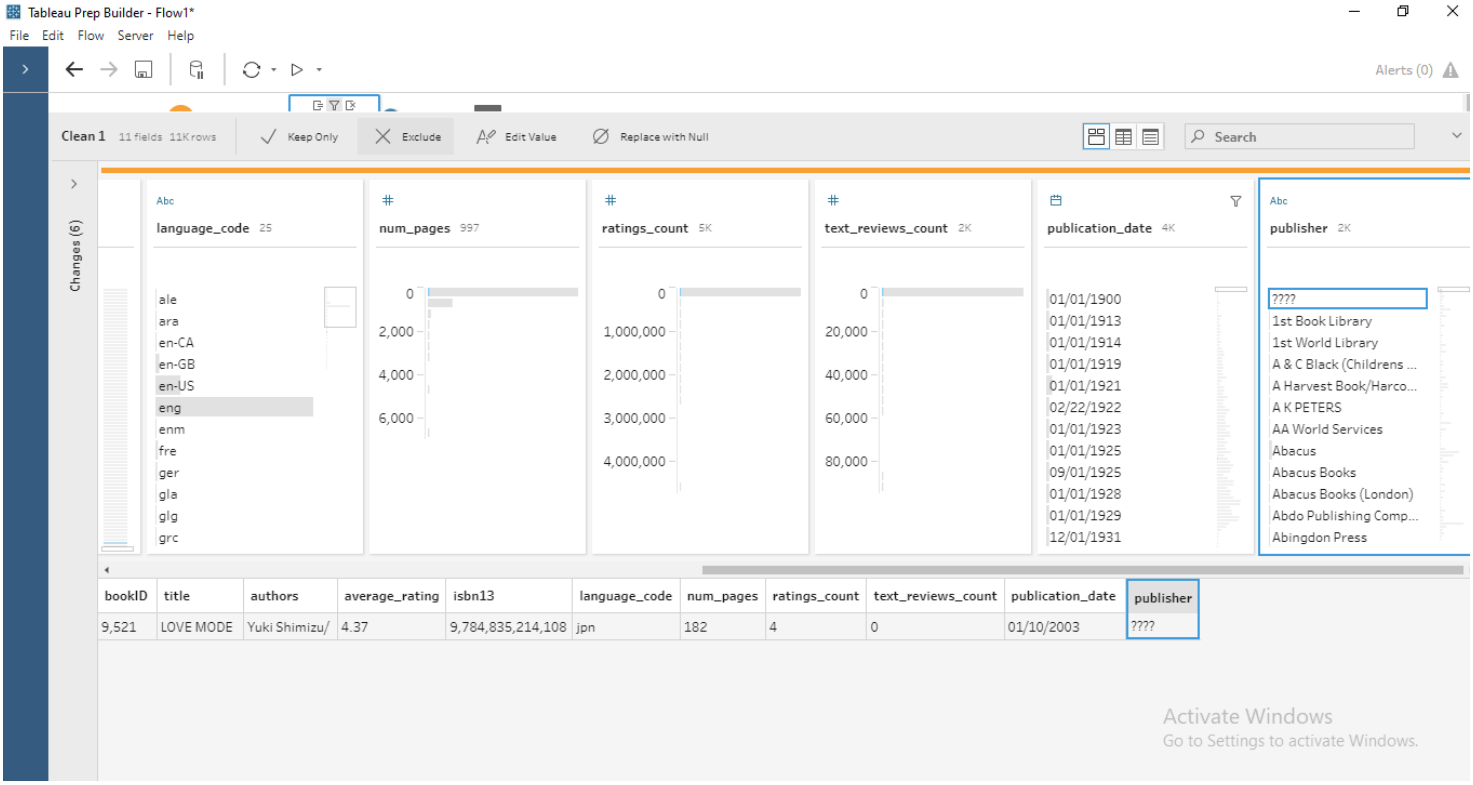
After removing null values:



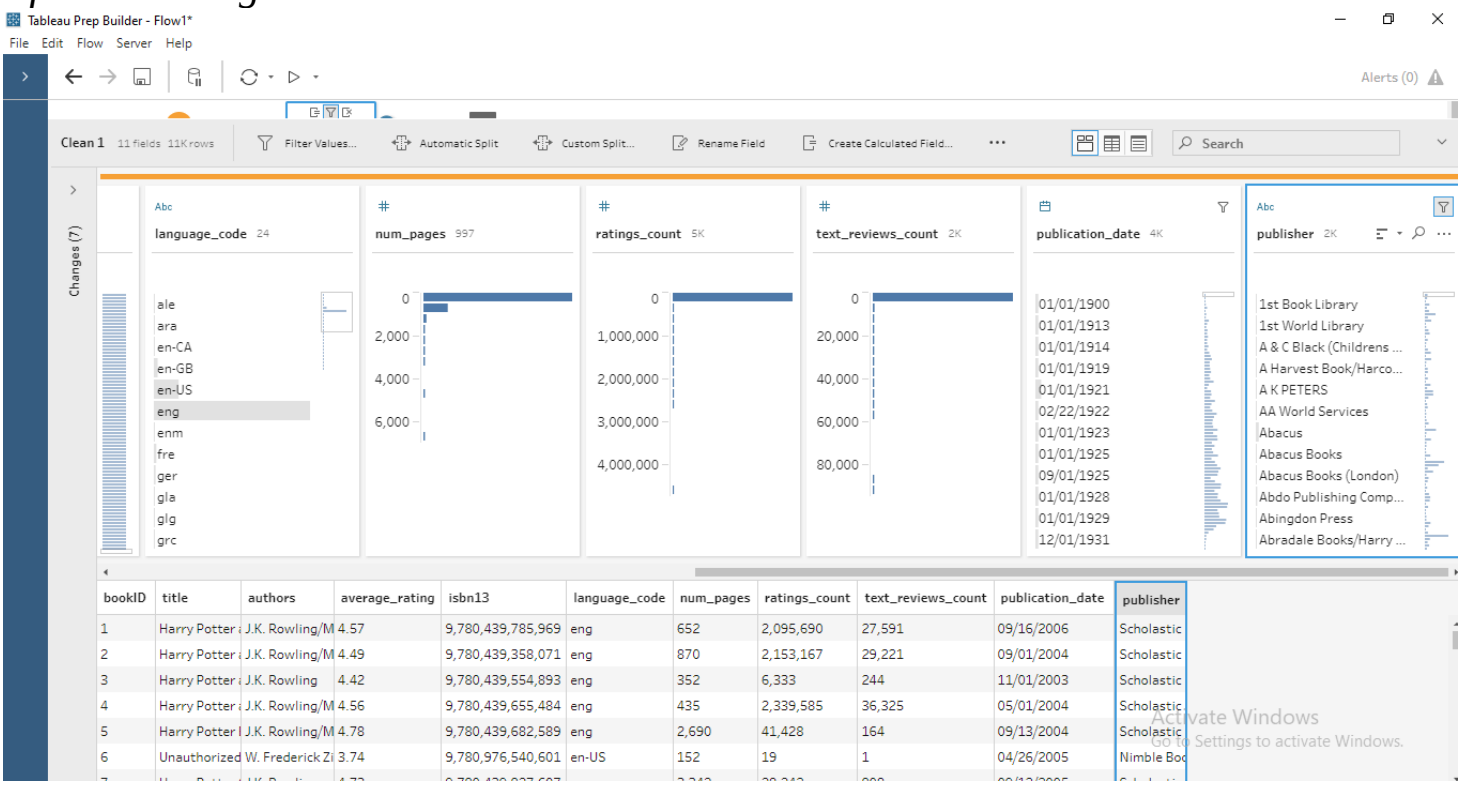
bookID	title	authors	average_rating	isbn13	language_code	num_pages	ratings_count	text_reviews_count
1	Harry Potter and the HalfBlood Prince Harry Pot	J.K. Rowling/Mary GrandPré	4.57	9,780,439,785,969	eng	652	2,095,690	27,591
2	Harry Potter and the Order of the Phoenix Harry	J.K. Rowling/Mary GrandPré	4.49	9,780,439,358,071	eng	870	2,153,167	29,221
3	Harry Potter and the Chamber of Secrets Harry F	J.K. Rowling	4.42	9,780,439,554,893	eng	352	6,333	244
4	Harry Potter and the Prisoner of Azkaban Harry	J.K. Rowling/Mary GrandPré	4.56	9,780,439,655,484	eng	435	2,339,585	36,325
5	Harry Potter Boxed Set Books Harry Potter	J.K. Rowling/Mary GrandPré	4.78	9,780,439,682,589	eng	2,690	41,428	164
6	Unauthorized Harry Potter Book Seven News Ha	W. Frederick Zimmerman	3.74	9,780,976,540,601	en-US	152	19	1

Activate Windows
Go to Settings to activate Windows.

Step-7: *Removing special characters from publisher*
Selecting by manual and exclude:



After removing:



Finally, after cleaning the data, the final dataset sample is shown below:

	B	C	D	E	F	G	H	I	J	K
1	title	authors	average_rating	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
2	A Backward Glance	Edith Wharton/Louis Auchinc	3.76	9780684847559	eng	385	521	77	15-Jul-1998	Scribner
3	A Beautiful Mind	Sylvia Nasar	4.13	9780571212927	eng	461	115201	1008	4-Feb-2002	Faber Faber
4	A Bend in the River	V.S. Naipaul	3.77	9780330487146	eng	326	13147	690	10-May-2002	Picador USA
5	A Bend in the Road	Nicholas Sparks	4.03	9780446696135	eng	341	125693	3229	1-Apr-2005	Grand Central Publishing
6	A Birthday for Frances	Russell Hoban/Lillian Hoban	4.24	9780099432449	eng	32	2256	71	6-Jun-2002	Harper Collins
7	A Bite to Remember Argeneau	Lynsay Sands	4.24	9780060774073	eng	362	17710	393	27-Jun-2006	Avon
8	A Book of Blue Flowers	Robert L. Geneve	3.83	9780881927696	eng	327	0	0	1-Feb-2006	Timber Press (OR)
9	A Book of Common Prayer	Joan Didion	3.8	9780679754862	eng	272	2962	255	11-Apr-1995	Vintage International
10	A Box of Matches	Nicholson Baker	3.67	9780375706035	eng	192	1679	192	9-Mar-2004	Vintage
11	A Break with Charity A Story about the Salem Witch Trials	Ann Rinaldi	3.79	9780152046828	eng	320	5052	417	1-Jul-2003	HMH Books for Young Readers
12	A Break With Charity A Story of the Salem Witch Trials	Ann Rinaldi	3.79	9780785735489	eng	297	37	6	1-Apr-1994	Turtleback Books
13	A Breath of Snow and Ashes Outlander	Diana Gabaldon	4.44	9780385340397	eng	980	105950	3992	29-Aug-2006	Delta
14	A Brief History of Time	Stephen Hawking	4.17	9780553380163	eng	212	239652	5860	1-Sep-1998	Bantam Books
15	A Briefer History of Time	Stephen Hawking/Leonard M	4.22	9780553804362	eng	176	24035	828	27-Sep-2005	Bantam
16	A Calendar of Wisdom Daily Thoughts to Nourish the Soul	Leo Tolstoy/Peter Sekirin	4.07	9780684837932	eng	384	502	43	14-Oct-1997	Scribner
17	A Caress of Twilight Merry Gentry	Laurell K. Hamilton	4.06	9780345478160	en-US	348	32029	596	15-Mar-2005	Ballantine Books
18	A Cargo of Women Susannah Watson and the Convicts of the Princess Royal	Babette Smith	3.58	9781877058332	eng	264	0	0	1-Jun-2005	Rosenberg Publishing
19	A Caribbean Mystery	Agatha Christie/Rosalind Ayr	3.8	9781572705494	eng	0	49	10	9-Oct-2006	Audiogo
20	A Caribbean Mystery Miss Marple	Agatha Christie	3.8	9780451199928	eng	224	18590	770	1-Mar-2000	Signet
21	A Carl Hiaasen Collection Stormy Weather Tourist Season and Strip Tease	Carl Hiaasen/Edward Asner	4.45	9780375410468	eng	9	44	1	4-Jan-2000	Random House Audio
22	A Case of Need	Jeffery Hudson/Michael Crich	3.72	9780451210630	eng	416	17135	654	5-Aug-2003	Signet
23	A Cat a Man and Two Women	Jun'ichiro Tanizaki/Paul McCa	3.8	9784770016058	eng	164	972	74	1-Apr-1992	Kodansha
24	A Cavern of Black Ice Sword of Shadows	J.V. Jones	3.87	9780765345516	eng	769	7919	163	1-Mar-2005	Tor Fantasy
25	A Certain Justice Abe Glitsky	John Lescroart	4.01	9780451217769	eng	624	2993	107	7-Feb-2006	Signet
26	A Certain Justice Adam Dalgliesh	P.D. James	3.93	9780770429911	eng	688	6877	374	28-Mar-2006	Seal Books
27	A Changeling for All Seasons Changeling Seasons	Angela Knight/Sahara Kelly/J	3.76	9781595962805	eng	304	167	4	1-Nov-2005	Changeling Press
28	A Childs Garden of Verses	Robert Louis Stevenson/Tash	4.3	9780689823824	eng	67	22227	431	1-Feb-1999	Simon & Schuster Books for Young R
29	A Chosen Few The Resurrection of European Jewry Readers Circle	Mark Kurlansky	3.87	9780345448149	eng	456	51	2	26-Mar-2002	Ballantine Books
30	A Christmas Carol and Other Christmas Books	Charles Dickens/Robert Doug	3.96	9780192806949	eng	438	156	20	1-Nov-2007	Oxford University Press USA
31	A Christmas Carol and Other Christmas Writings	Charles Dickens/Michael Slat	4.14	9780140439052	eng	288	30393	449	30-Oct-2003	Penguin Classics
32	A Christmas Carol and The Night Before Christmas	Charles Dickens/Clement C. M	4.35	9780517229279	eng	156	91	15	3-Oct-2006	Gramercy Books
33	A Christmas Carol Great Illustrated Classics	Malvina G. Vogel/Charles Dic	4.36	9781596792371	eng	238	466	31	1-Jan-2005	Abdo Publishing Company
34	A Christmas Memory	Truman Capote	4.24	9780375837890	eng	48	9982	849	10-Oct-2006	Knopf Books for Young Readers
35	A Christmas Story	Jean Shepherd	4.08	9780767916226	eng	144	3187	381	28-Oct-2003	Broadway Books
36	A Civil Action	Jonathan Harr	3.97	9780679772675	eng	512	13628	754	27-Aug-1996	Vintage
37	A Clash of Kings A Song of Ice and Fire	George R.R. Martin	4.44	9780553381696	eng	969	538766	16535	28-May-2002	Bantam

Plot Area

Analysis on dataset

1. Total books published in each year

Introduction

- *By performing this analysis, we will get number of books published in each year*

Description:

It is to see the number of books published in each year.

Specific requirements, functions, and formulas:

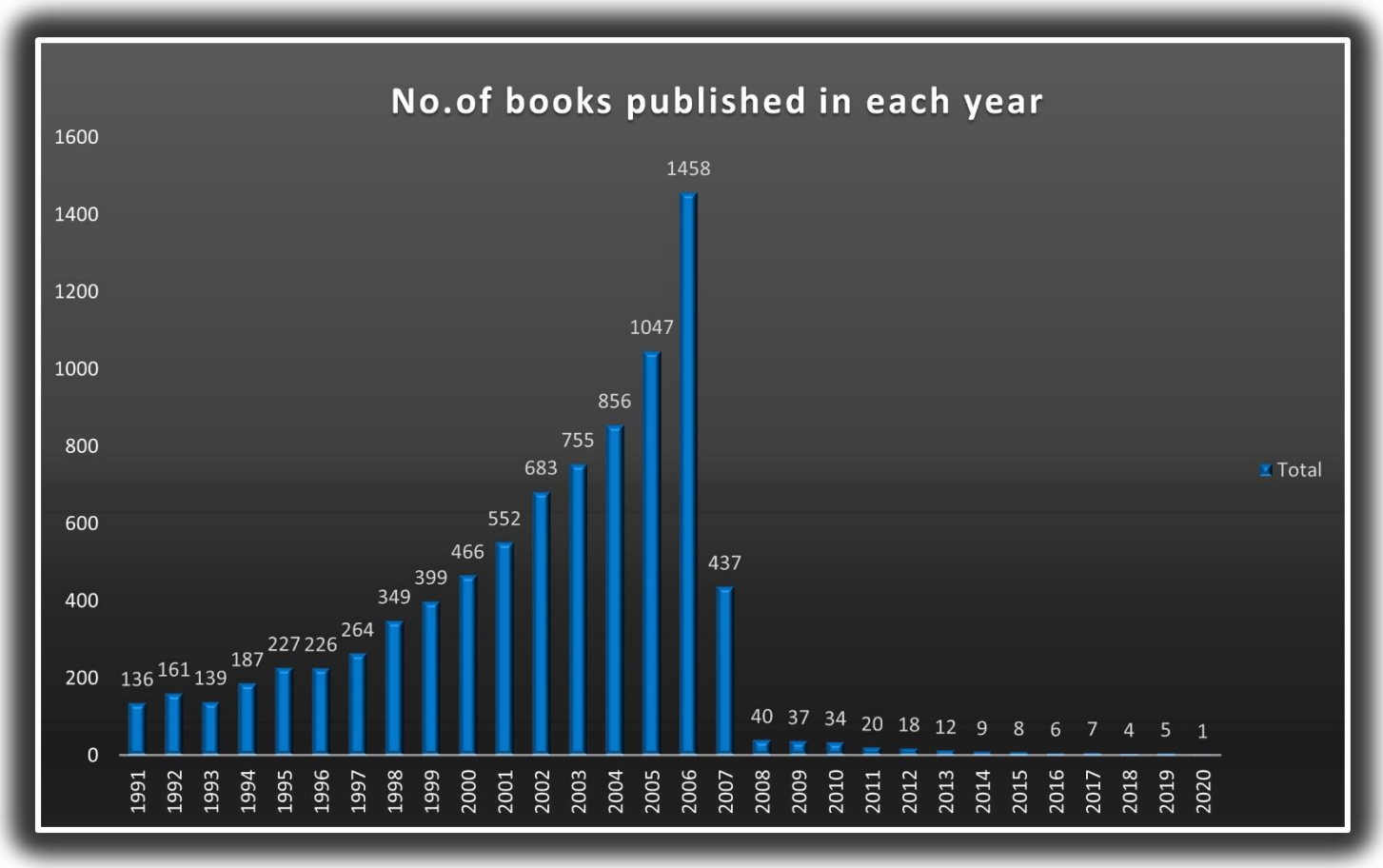
For counting number of books we can use count function: =count()

Analysis results: *In the year 2006, most number of books published i.e., 1458.*

Year	No.of books
1991	136
1992	161
1993	139
1994	187
1995	227
1996	226
1997	264
1998	349
1999	399
2000	466
2001	552
2002	683
2003	755
2004	856
2005	1047
2006	1458
2007	437
2008	40
2009	37
2010	34
2011	20
2012	18
2013	12
2014	9
2015	8
2016	6
2017	7
2018	4
2019	5
2020	1
Grand Total	8543

Slicer:

Visualization:



2. Top 50 average rated books

Introduction

- *By performing this analysis, we will get the top 50 books with respect to average rating*

Description:

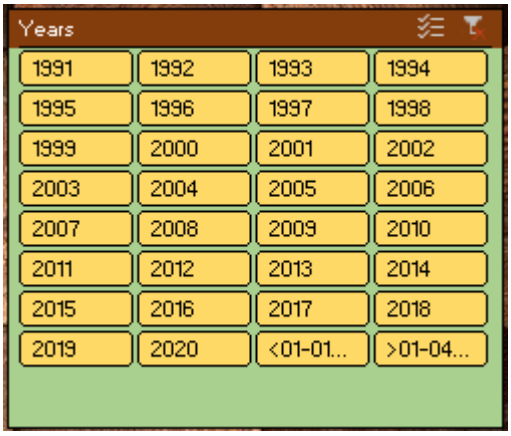
The analysis is based on top 50 rated books in the dataset with respect to average_rating

Analysis results:

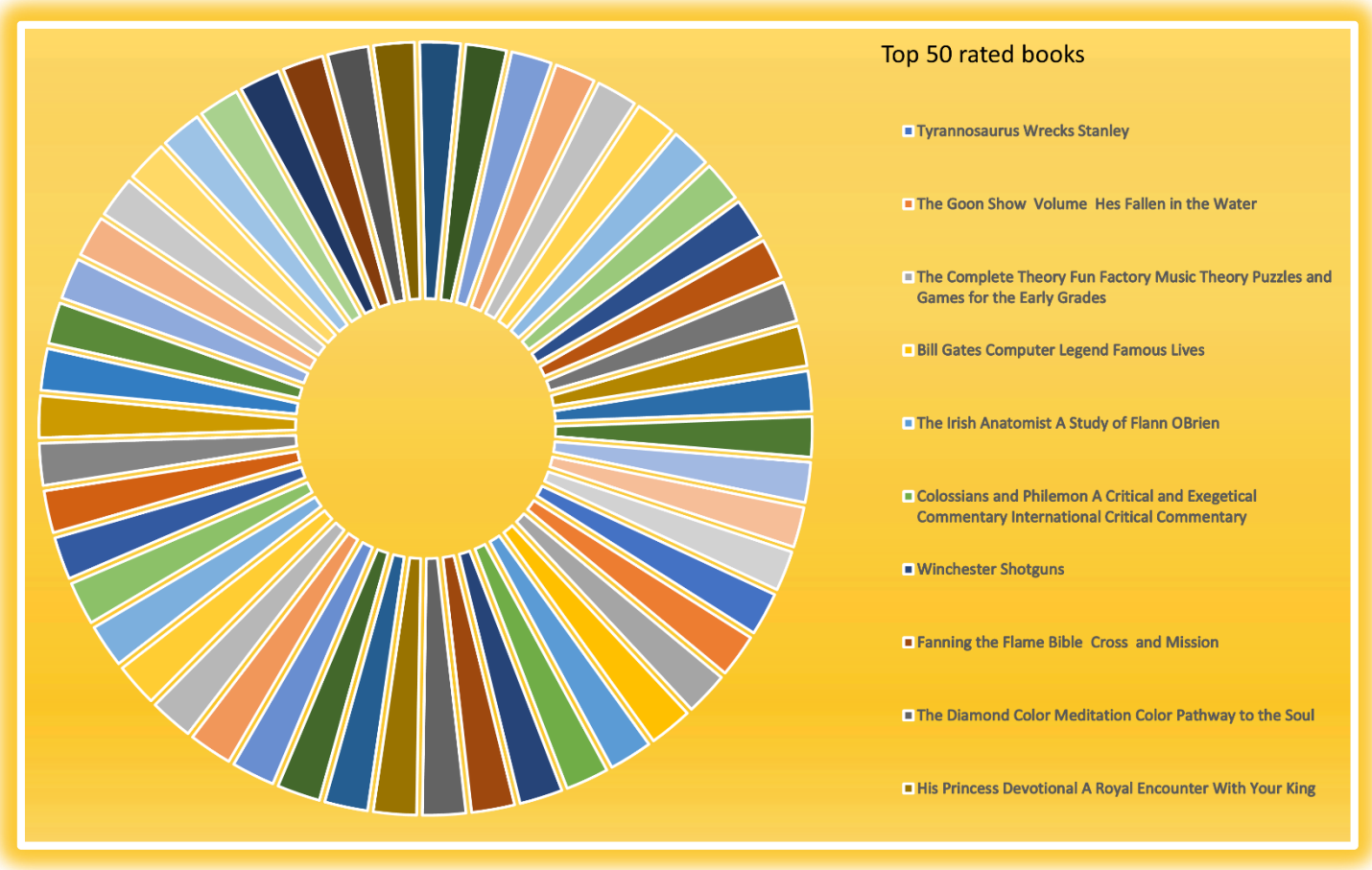
This analysis shows the top 50 books with top rating and names of books.

Title	average rating
Tyrannosaurus Wrecks Stanley	5
The Goon Show Volume Hes Fallen ir	5
The Complete Theory Fun Factory Mu	5
Bill Gates Computer Legend Famous L	5
The Irish Anatomist A Study of Flann C	5
Colossians and Philemon A Critical an	5
Winchester Shotguns	5
Fanning the Flame Bible Cross and M	5
The Diamond Color Meditation Color P	5
His Princess Devotional A Royal Enco	5
The Goon Show Volume My Knees H	5
Literature Circle Guide Bridge to Terab	5
The New Big Book of America	5
Middlesex Borough Images of America	5
Willem de Kooning Late Paintings	5
Zone of the Enders The nd Runner Off	5
The American Campaign US Presider	5
Oliver Wendell Holmes in Paris Medicir	5
Existential Meditation	4.91
Little Big Book for Gods Children	4.88
Stargirl LitPlans on CD	4.86
Corvette	4.83
The Complete Calvin and Hobbes	4.82
Harry Potter Boxed Set Books Harry P	4.78
Wissenschaft der Logik Die Lehre Von	4.78
Its a Magical World Calvin and Hobbes	4.76
True Harvest Readings From Henry Da	4.75
The John Deere TwoCylinder Tractor E	4.75
Harry Potter Collection Harry Potter	4.73
Early Color	4.73
Homicidal Psycho Jungle Cat Calvin a	4.72
The Beatles Story on Capitol Records	4.72
Elliott Erwitt Snaps	4.72

Slicer:



Visualization:



3. Most rating counts containing books in each year

Introduction

- *By performing this analysis, we will get maximum rating counts book in a particular year*

Description:

The analysis based on maximum of ratings count containing books in each year

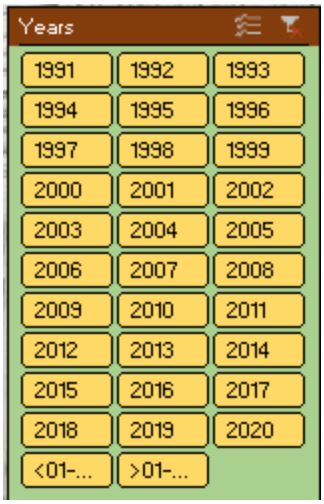
Specific requirements, functions, and formulas:

Select the cell below the given Quantity and apply the formula '=max ()'. This function will give the maximum number in a range of cells

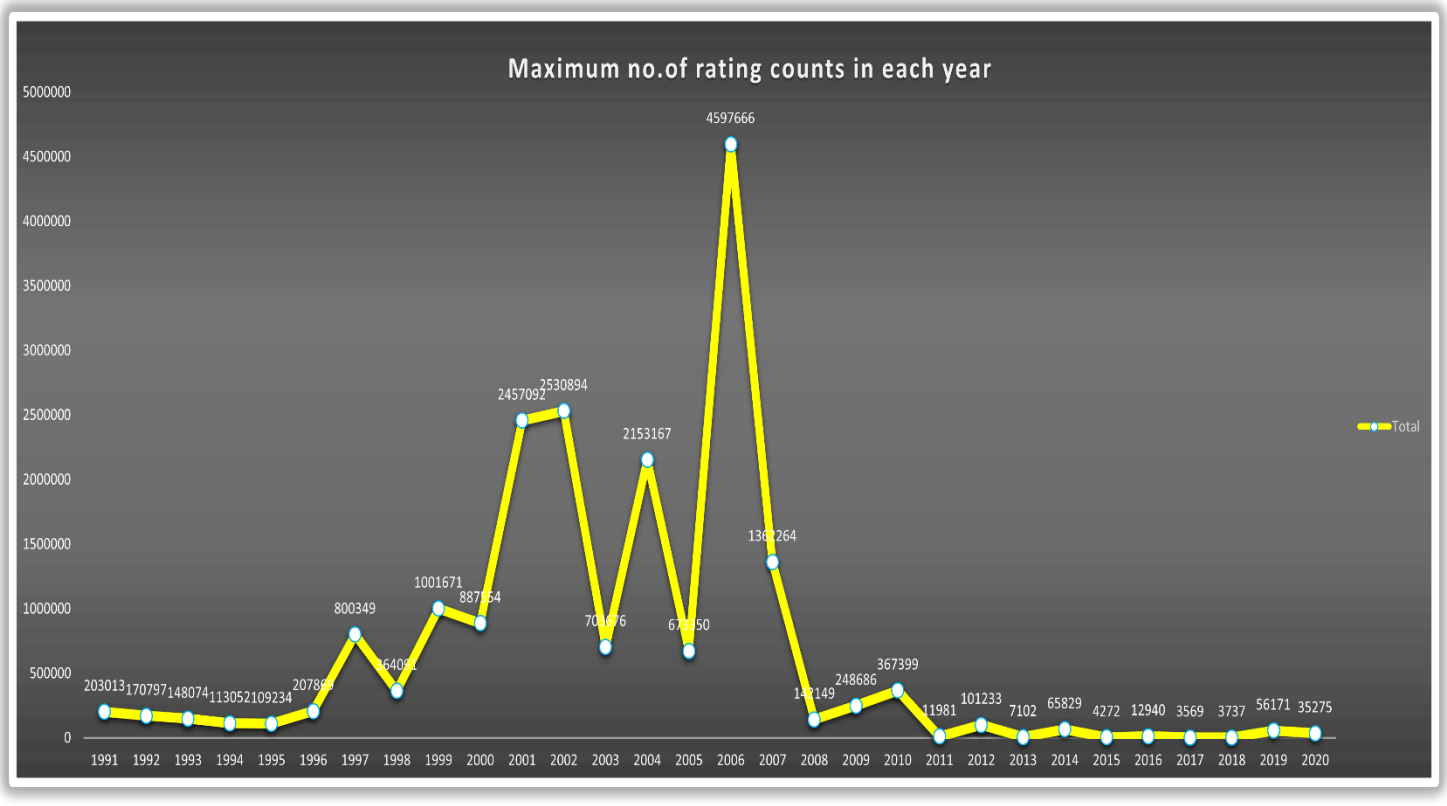
Analysis results:

Year	Max of ratings_count
1991	203013
1992	170797
1993	148074
1994	113052
1995	109234
1996	207869
1997	800349
1998	364091
1999	1001671
2000	887554
2001	2457092
2002	2530894
2003	703676
2004	2153167
2005	673350
2006	4597666
2007	1362264
2008	142149
2009	248686
2010	367399
2011	11981
2012	101233
2013	7102
2014	65829
2015	4272
2016	12940
2017	3569
2018	3737
2019	56171
2020	35275
Grand Total	4597666

Slicer:



Visualization:



4. Number of books written in each language

Introduction

- *By performing this analysis, we will get number of books written in each language with respect to language_code*

Description:

In this analysis, we display the language code and number of books written in that specific language.

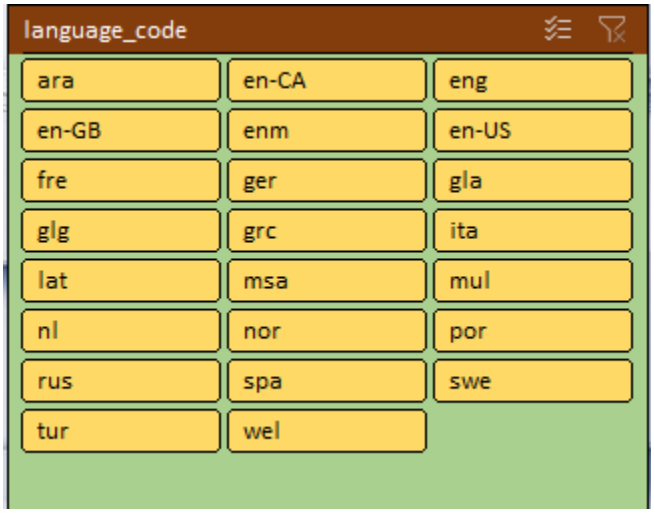
Specific requirements, functions, and formulas:

Select the cell below the given Quantity and apply the formula '=count()'. This function will count the number of books in a range of cells

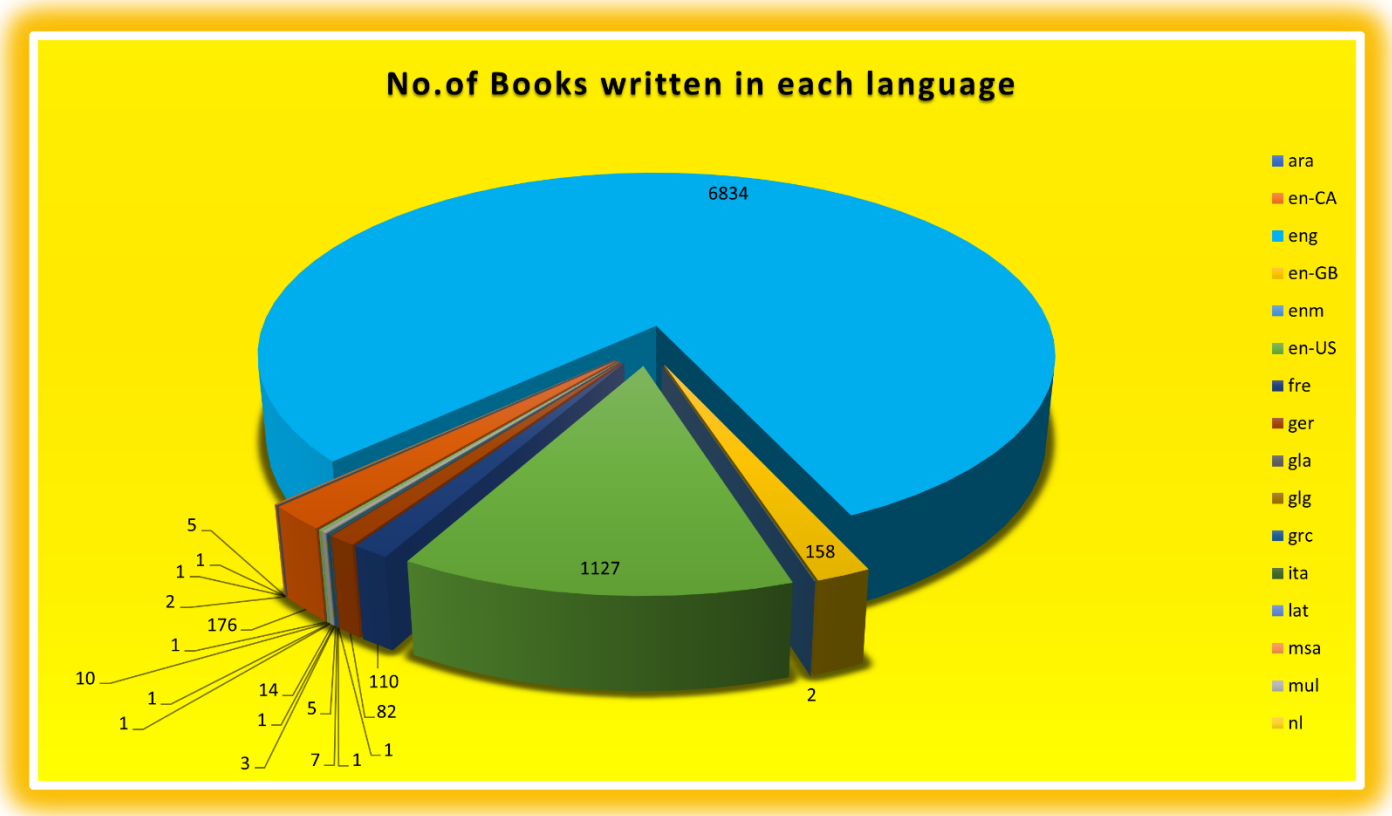
Analysis results:

Language code	No.of books
ara	1
en-CA	5
eng	6834
en-GB	158
enm	2
en-US	1127
fre	110
ger	82
gla	1
glg	1
grc	7
ita	5
lat	3
msa	1
mul	14
nl	1
nor	1
por	10
rus	1
spa	176
swe	2
tur	1
Grand Total	8543

Slicer:



Visualization:



5. *Top 10 text reviewed books*

Introduction

❖ *By performing this analysis, we will get top 10 text reviewed books*

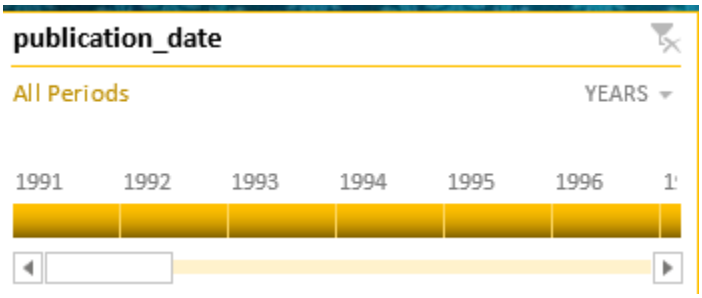
Description:

10 Most text reviewed books gives which are the best books in the whole dataset.

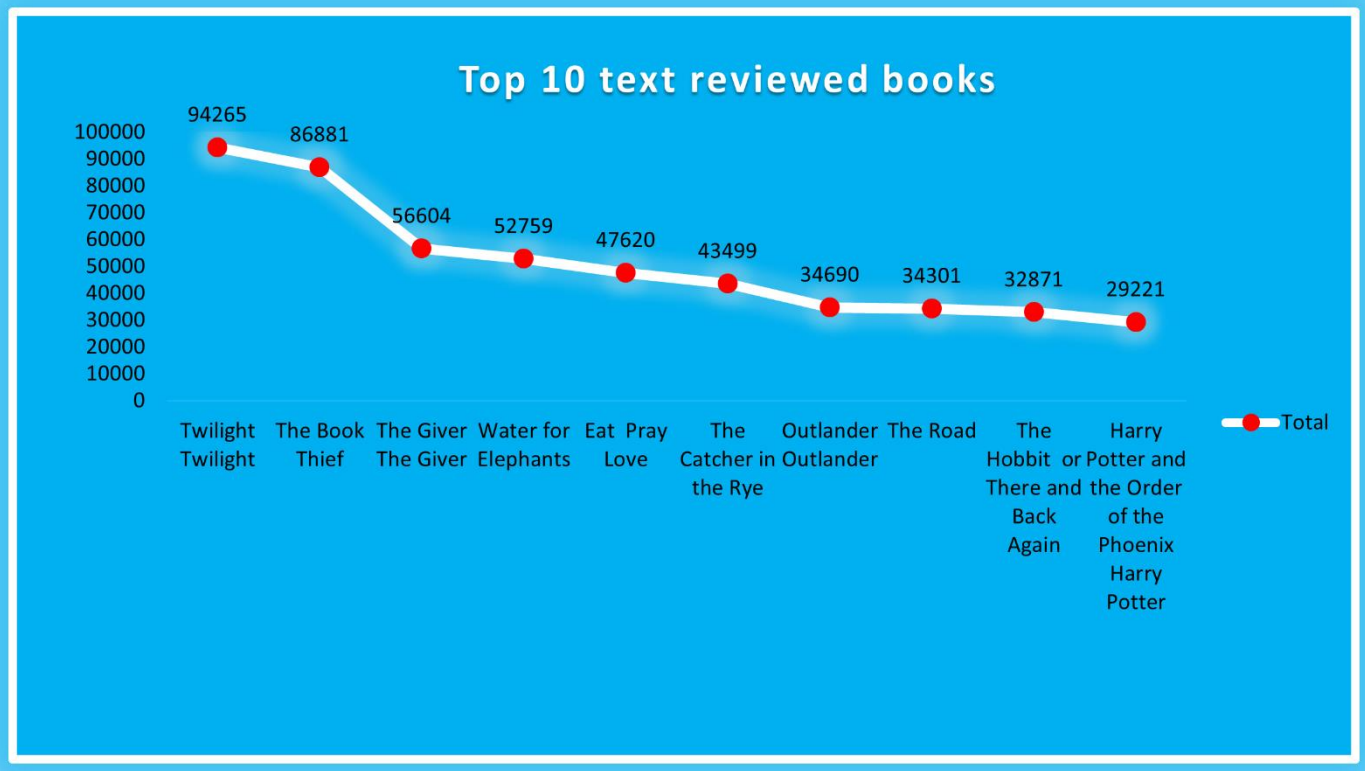
Analysis results:

Book Title	Text Reviews Count
Twilight Twilight	94265
The Book Thief	86881
The Giver The Giver	56604
Water for Elephants	52759
Eat Pray Love	47620
The Catcher in the Rye	43499
Outlander Outlander	34690
The Road	34301
The Hobbit or There and Back Again	32871
Harry Potter and the Order of the Phoenix Harry Potter	29221
Grand Total	**512711**

Time line:

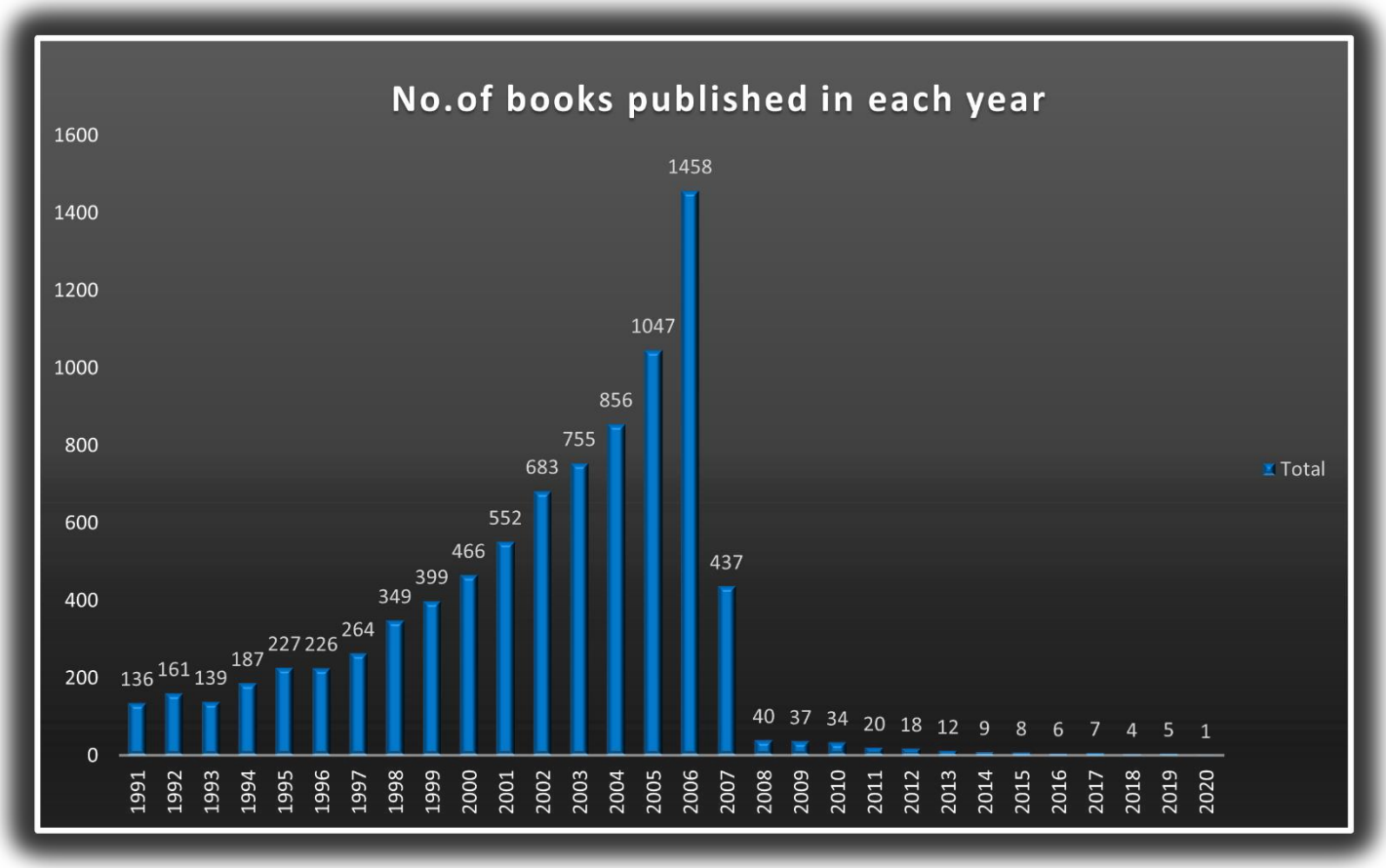


Visualization:

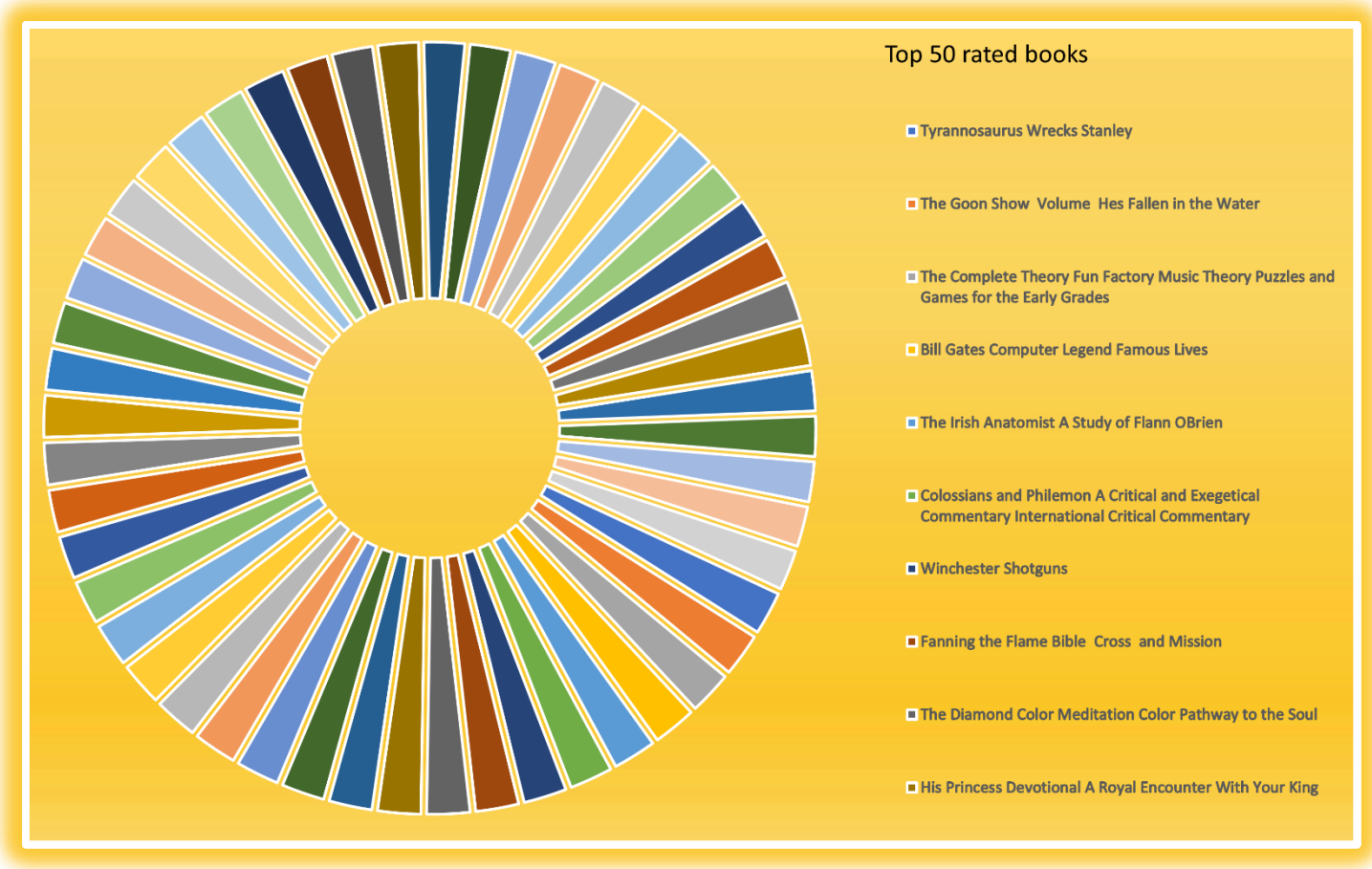


List of Analysis with results

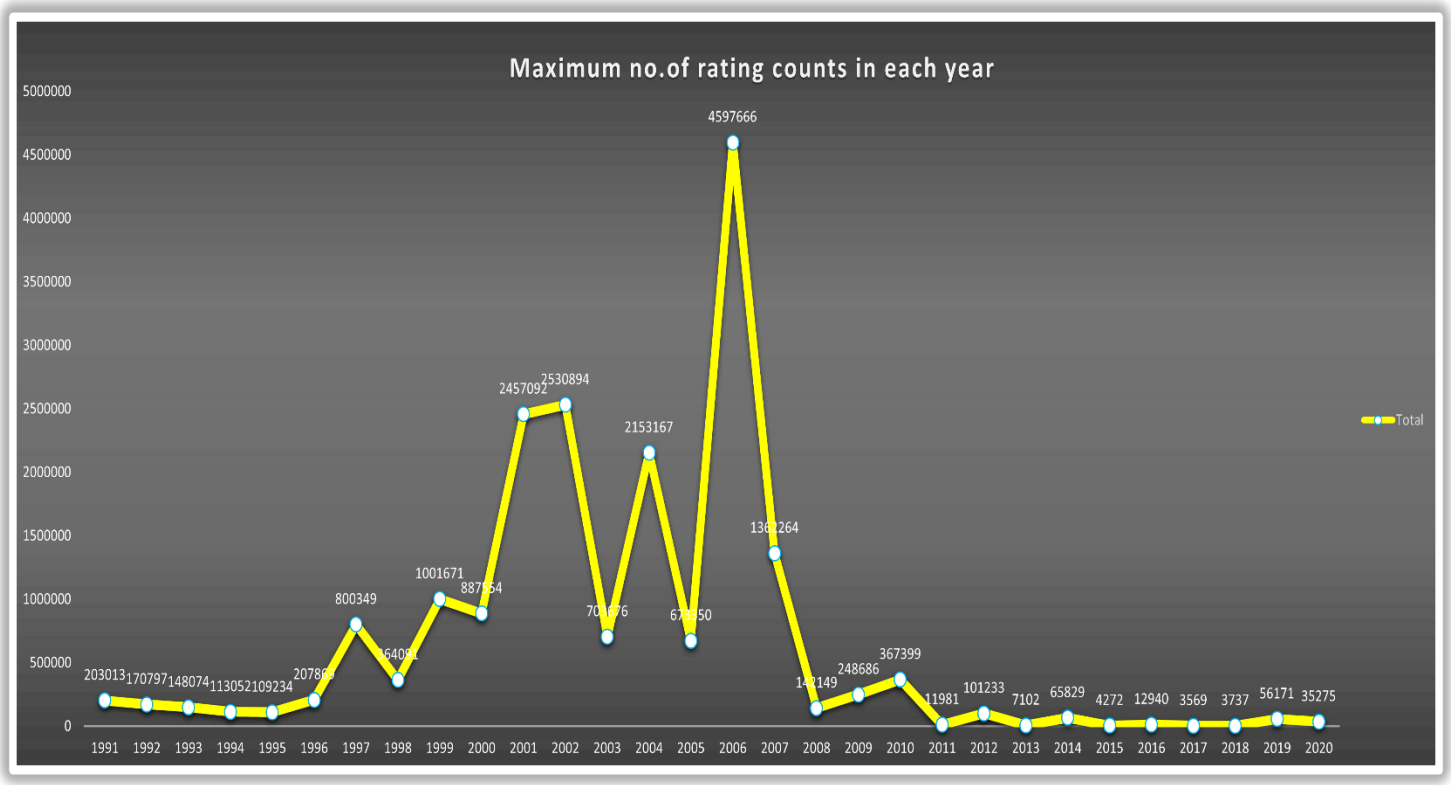
1. Total books published in each year:



2. Top 50 rated books:

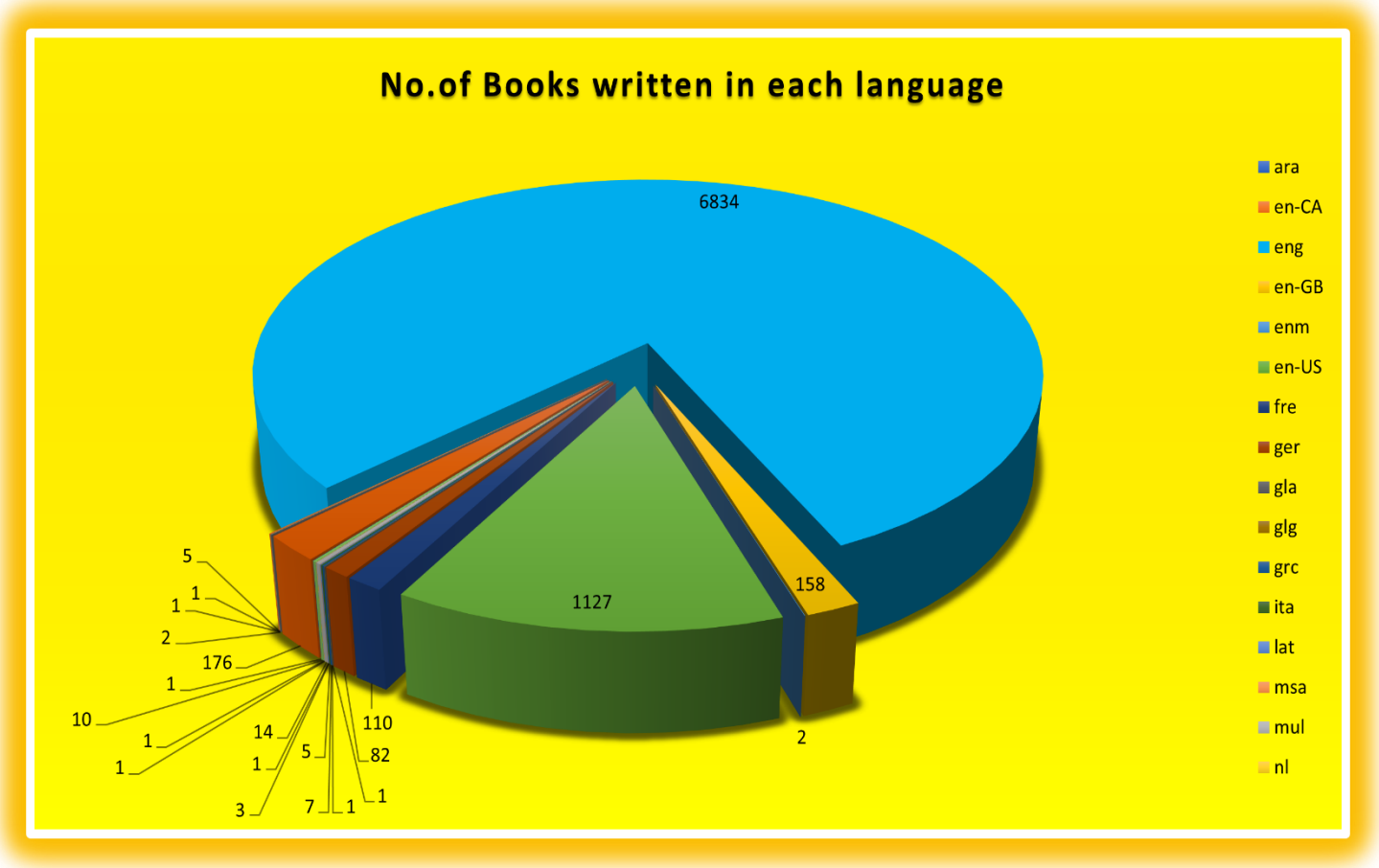


3. Most rating counts containing books in each year:

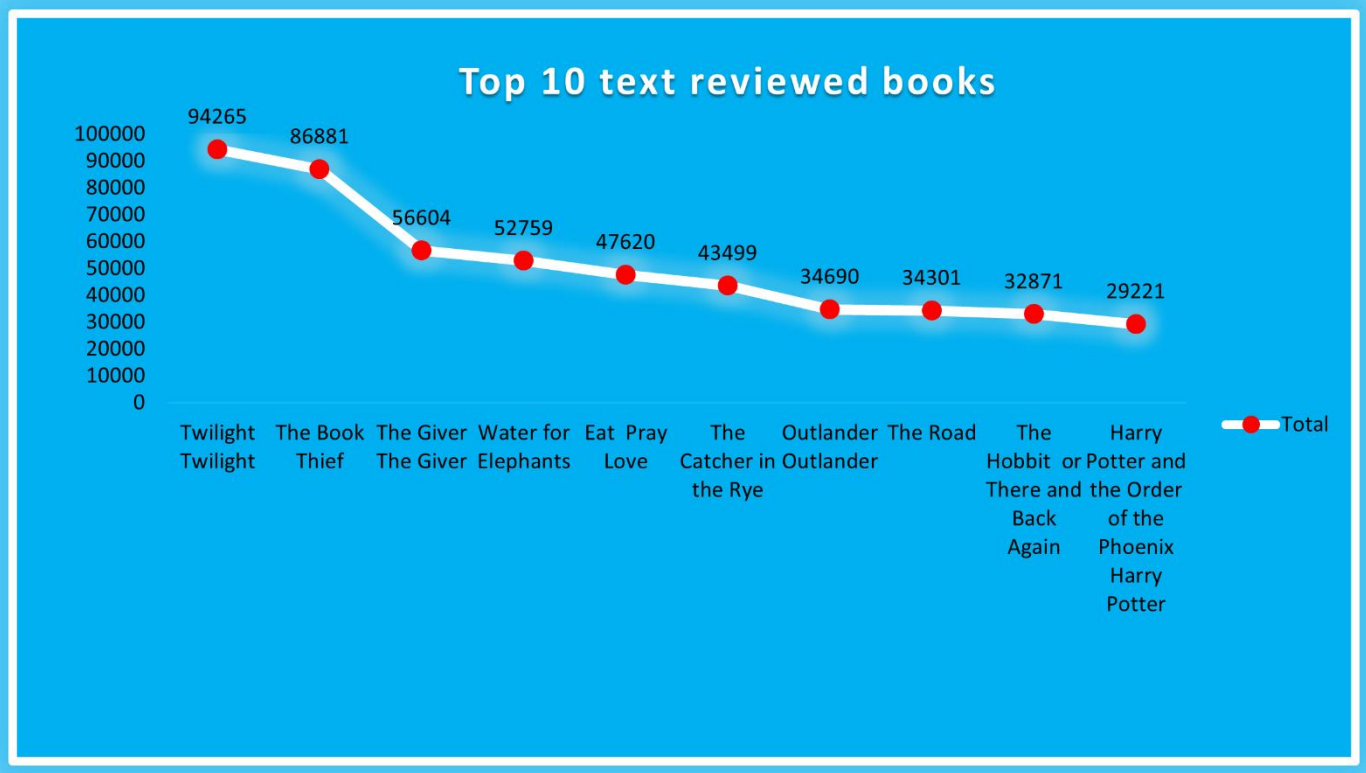


.

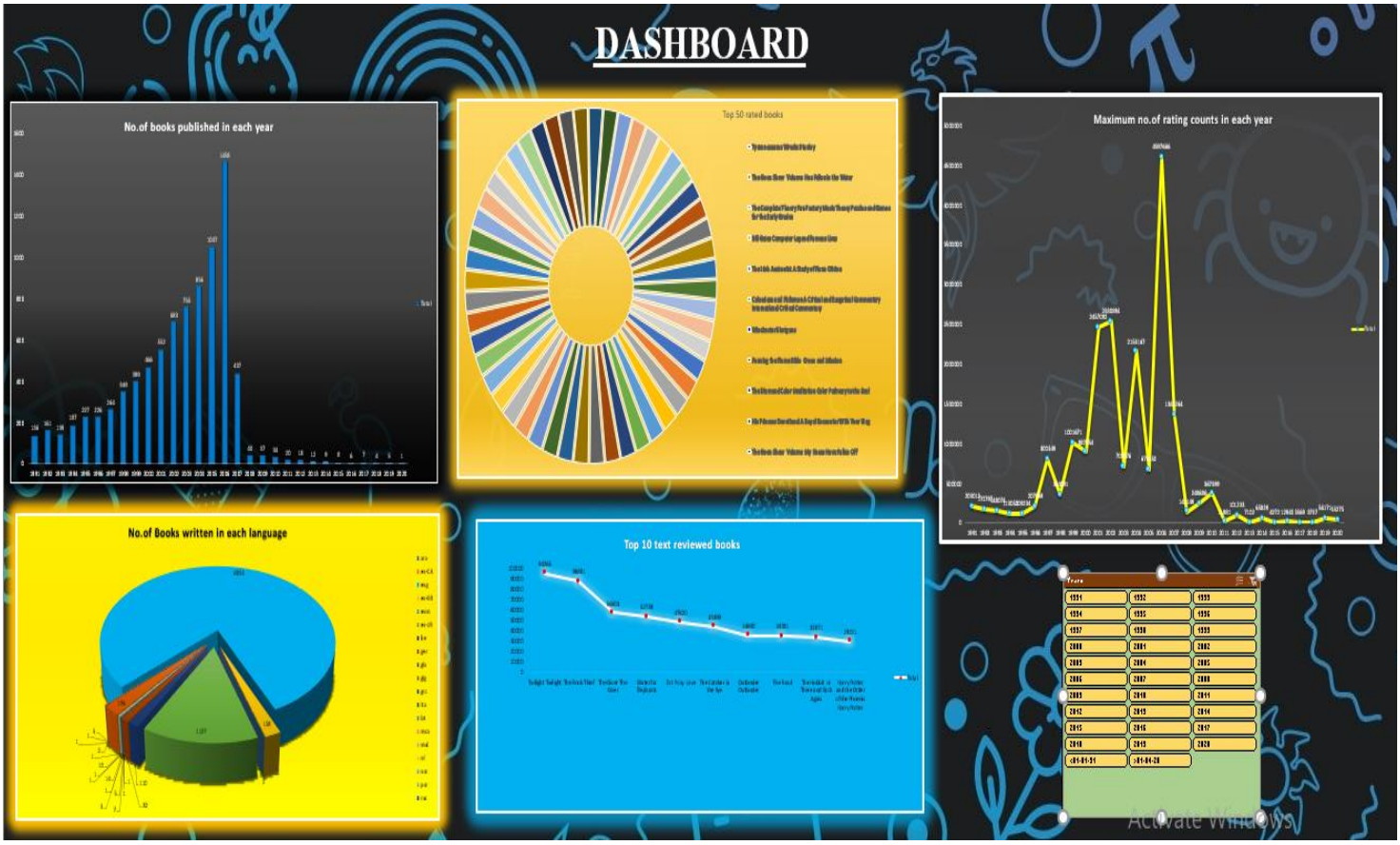
4. No. of books written in each language:



5. Top 10 text reviewed books:



FINAL DASHBOARD:



BIBLIOGRAPHY:

- *Kaggle*
- *Trending videos of YouTube analysis*
- *https://www.kaggle.com/datasnaek/youtube-new*
- *https://www.analyticsvidhya.com/blog/2019/09/7-data-science- projects-github-showcase-your-skills/*
- *https://en.wikipedia.org/wiki/YouTube*
- *https://www.quora.com/How-does-Youtubes-algorithm-works-in-terms-of-trending-a-video*